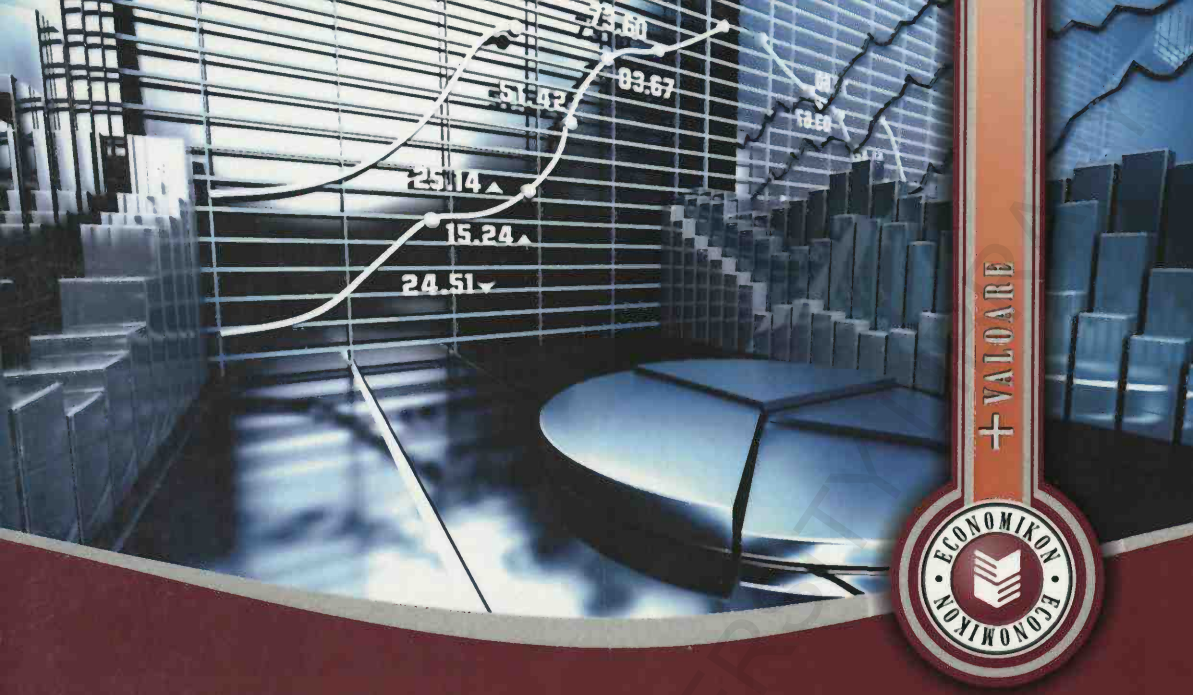


Econometrie

cu aplicații în R

Ediția a IV-a
revăzută
și adăugită

DĂNUȚ-VASILE JEMNA

III323 154
BCU IASI

Editura Universității „Alexandru Ioan Cuza" Iași

III323 154
BCU IASI

BCU IASI/CENTRAL UNIVERSITY LIBRARY

498.

BCU IASI/CENTRAL UNIVERSITY LIBRARY

Colecția **Economikon** este coordonată de prof.univ.dr. Ion Pohoață

Colecție fondată de prof.univ.dr. Ion Pohoață și prof.univ.dr. Dumitru Zaiț

Dănuț-Vasile Jemna este profesor universitar la Facultatea de Economie și Administrarea Afacerilor, Universitatea „Al.I. Cuza" din Iași. În cadrul specializării de Statistică și Previziune Economică, predă cursurile de Teoria și practica sondajelor, Econometrie și Demografie, la nivel de licență, iar la programele de master predă Econometrie avansată, Metodologia cercetării și Statistică bayesiană. A efectuat stagii de cercetare doctorală și postdoctorală la Universitatea Poitiers din Franța și a realizat numeroase mobilități de predare în cadrul programelor Erasmus în țări precum Italia, Franța, Polonia.

Referent științific:
Prof.univ.dr. Mihai Roman

Redactor: Marius-Nicușor Grigore
Coperta: Manuela Oboroceanu
Tehnoredactor: Florentina Crucerescu

ISBN: 978-606-714-384-3

© 2017, Editura Universităţii „Alexandru Ioan Cuza"
700109 – Iaşi, str. Pinului, nr. 1A, tel./fax: (0232) 314947
http:// www.editura.uaic.ro e-mail: editura@uaic.ro

BCU - CENTRAL UNIVERSITY LIBRARY

DĂNUȚ-VASILE JEMNA

782762
336.454

Econometrie

cu aplicații în R

Ediția a IV-a, revăzută și adăugită

ALE

0 000010 71092 BCU IASI

Editura Universităţii „Alexandru Ioan Cuza" din Iaşi

2017

BCU IASI SICENTRAL UNIVERSITY LIBRARY

Descrierea CIP a Bibliotecii Naţionale a României
JEMNA, DĂNUŢ
Econometrie: cu aplicaţii în R / Dănuţ-Vasile Jemna. - Ed. a 4-a, rev. şi adăug.. - Iaşi: Editura Universităţii „Al. I. Cuza", 2017
ISBN 978-606-714-384-3

330

15. FEB. 2018

BCU IAŞI CENTRAL UNIVERSITY LIBRARY

Cuprins

INTRODUCERE 9

1. ELEMENTE TEORETICE ŞI METODOLOGICE 11
1.1. Provocări pentru cercetare 12
1.2. Ce este econometria? 16
1.3. Probleme teoretice 19
1.3.1. Definirea conceptului de model econometric 19
1.3.2. Noţiuni, termeni şi notaţii 21
1.4. Probleme metodologice 24
1.4.1. Demersul metodologic al cercetării econometrice 24
1.4.2. Tipuri de date statistice utilizate în econometrie 26
1.5. Introducere în R 27
1.5.1. Mediul de lucru 27
1.5.2. Analiza statistică în R 34
1.6. Teste şi exerciţii 41

2. ELEMENTE DE TEORIA PROBABILITĂŢILOR ŞI DE STATISTICĂ MATEMATICĂ 45
2.1. Conceptul de probabilitate 45
2.1.1. Experienţă aleatoare. Eveniment aleator 45
2.1.2. Definiţia probabilităţii 47
2.2. Variabile aleatoare 49
2.2.1. Definiţie şi tipuri de variabile aleatoare 50
2.2.2. Caracteristici numerice ale unei variabile aleatoare 52
2.3. Parametri. Estimatori. Estimaţii 55
2.4. Repartiţii probabiliste bidimensionale 56
2.5. Repartiţii probabiliste frecvent utilizate în econometrie 59
2.6. Statistici pentru eşantioane extrase din colectivităţi cu repartiţie normală 69

BCU CENTRAL UNIVERSITY LIBRARY

2.7. Distribuții probabiliste în R 71
2.8. Teste și exerciții 74

3. REGRESIA LINIARĂ SIMPLĂ 79
3.1. Noțiuni introductive 79
3.1.1. Istoric 79
3.1.2. Tipuri de regresie 80
3.1.3. Interpretarea geometrică și statistică a regresiei 82
3.2. Modelul econometric de regresie liniară simplă 84
3.2.1. Prezentarea modelului 84
3.2.2. Ipotezele clasice ale modelului de regresie 88
3.3. Estimarea parametrilor modelului 89
3.3.1. Probleme ale estimării parametrilor modelului de regresie 90
3.3.2. Estimarea punctuală și prin interval de încredere 96
3.3.3. Regresia liniară simplă în R 99
3.4. Indicatori de corelație 105
3.4.1. Coeficientul de corelație 105
3.4.2. Raportul de determinație și raportul de corelație 108
3.5. Testarea parametrilor și a modelului de regresie 111
3.5.1. Etapele testării 111
3.5.2. Testarea parametrilor modelului 113
3.5.3. Testarea modelului de regresie 116
3.5.4. Testarea indicatorilor de corelație 119
3.6. Probleme specifice modelului de regresie liniară simplă 122
3.6.1. Regresia prin origine 122
3.6.2. Liniaritatea și limitele modelului liniar simplu 124
3.7. Studiu de caz 125
3.8. Teste și exerciții 135

4. REGRESIA LINIARĂ MULTIPLĂ 141
4.1. Prezentarea modelului 141
4.2. Estimarea parametrilor modelului 143
4.3. Indicatori de corelație 148
4.3.1. Coeficienți de corelație bivariată, parțială și semi-parțială 148
4.3.2. Raportul de determinație multiplă și raportul de corelație multiplă 151
4.3.3. Coeficientul de corelație multiplă 153
4.3.4. Raportul de determinație multiplă ajustat 154
4.4. Testarea parametrilor și a modelului 155
4.4.1. Testarea coeficienților de regresie 155
4.4.2. Testarea modelului de regresie 157
4.4.3. Testarea influenței marginale a unei variabile independente 159

BCU BUCUREȘTI LIBRARY

4.5. Probleme specifice modelelor de regresie liniară multiplă 162
4.5.1. Modele cu restricții 162
4.5.2. Modele cu variabile standardizate 166
4.6. Studiu de caz 167
4.7. Teste și exerciții 173

5. REGRESIA NELINIARĂ 179
5.1. Modelul log-liniar 180
5.1.1. Estimarea modelului 180
5.1.2. Elasticitatea 182
5.1.3. Funcția de producție (Funcția Cobb-Douglas) 185
5.2. Modele semi-logaritmice 188
5.2.1. Modele cu variabila dependentă logaritmată 188
5.2.2. Modele cu variabila independentă logaritmată 193
5.3. Modelul reciproc. Curba lui Philips 195
5.4. Modele polinomiale 199
5.4.1. Modelul parabolic sau quadratic 199
5.4.2. Modelul cubic 202
5.5. Teste și exerciții 204

6. MODELE DE REGRESIE CU VARIABILE ALTERNATIVE 211
6.1. Modele de regresie cu variabile alternative independente 211
6.1.1. Modele ANOVA 212
6.1.2. Modele ANCOVA 221
6.1.3. Aplicații ale modelelor ANOVA și ANCOVA 229
6.2. Modele de regresie cu variabile alternative dependente 237
6.2.1. Probleme ale estimării modelului cu variabilă alternativă dependentă 237
6.2.2. Modelul Logit 238
6.2.3. Modelul Probit 244
6.3. Teste și exerciții 246

7. VERIFICAREA IPOTEZELOR MODELULUI DE REGRESIE 253
7.1. Ipoteze asupra erorilor 254
7.1.1. Media variabilei reziduale este egală cu zero 254
7.1.2. Homoscedasticitatea erorilor 259
7.1.3. Normalitatea erorilor 270
7.1.4. Necorelarea erorilor 276
7.2. Ipoteze asupra variabilelor independente 285
7.2.1. Prezentare ipoteze 285
7.2.2. Ipoteza de necoliniaritate a variabilelor independente 285
7.3. Teste și exerciții 289

8. MODELAREA SERIILOR DE TIMP 295
8.1. Seriile de timp ca procese stocastice 295
8.1.1. De la abordarea statistică la abordarea econometrică a unei serii de timp 295
8.1.2. Procese stocastice. Definiţie şi caracteristici 297
8.1.3. Staţionaritatea proceselor stocastice. Procese integrate 301
8.2. Modelarea seriilor de timp cu ajutorul proceselor stocastice 311
8.2.1. Procese de tip medie mobilă (MA) 311
8.2.2. Procese autoregresive (AR) 313
8.2.3. Procese ARMA 316
8.2.4. Procese ARIMA. Metoda Box-Jenkins 317
8.3. Teste şi exerciţii 323

9. MODELE ECONOMETRICE DINAMICE 327
9.1. Specificul şi importanţa modelelor dinamice 327
9.2. Estimarea modelelor dinamice 329
9.2.1. Estimarea modelelor cu decalaj distribuit 329
9.2.2. Estimarea modelelor autoregresive 335
9.3. Teste şi exerciţii 338

BIBLIOGRAFIE 341

TABELE PROBABILISTE 347

BCU IASI/CENTRAL UNIVERSITY LIBRARY

B.C.U. „M. EMINESCU" IAȘI

Introducere

Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference.
P.A. Samuelson

Propunem o formulă a cărții de *Econometrie* în care aducem câteva modificări importante față de aparițiile anterioare. În primul rând, elementele aplicative, exemplele și exercițiile sunt realizate cu ajutorul pachetului program R. Această orientare este coroborată cu o tot mai largă utilizare a acestui software în companii și universități. În al doilea rând, pentru realizarea exemplelor s-a apelat, în general, la date statistice reale la nivelul României, Uniunii Europene sau la nivel mondial, cu scopul de a ajuta cititorul să vadă mai clar maniera în care se poate utiliza metoda econometrică pentru obiective practice concrete. În al treilea rând, am adus modificări la nivelul fiecărui capitol, fie prin anumite completări și clarificări metodologice, fie prin noi subiecte de interes pentru orice posibil utilizator al acestei lucrări. Nu în ultimul rând, fiecare capitol se încheie cu o secțiune de teste și exerciții utile pentru evaluarea conținutului însușit.

Preocupările de astăzi ale economiștilor sunt legate mai degrabă de construirea de modele decât de dezvoltarea de noi teorii. Orientarea practică îi determină să caute soluții la probleme concrete și să propună politici și măsuri de decizie economică având ca bază de plecare o cantitate imensă de date empirice, adică de măsurători ale fenomenelor economice și sociale.

Cu ajutorul computerului și al unor baze de date diversificate ca volum și structură, specialistul de astăzi creează modele, în condiții pe care părinții econometriei nu le-au putut prevedea. În ultima vreme, distanța dintre teoria și practica economică s-a redus tot mai mult. Specialistul din domeniul economic, fie că este la catedră sau într-un institut de cercetare, fie că lucrează într-o firmă de producție sau într-o instituție ce oferă servicii, este implicat într-un proces de formare continuă și este pus în situația de a integra cele două lumi (teoria și practica) care cândva erau mult prea îndepărtate și cu orizonturi adesea diferite.

Dacă la origini, modelarea fenomenelor economice (econometria) era parte integrantă a muncii de cercetare a economistului generic, astăzi asistăm la un puternic fenomen de specializare care face din econometrie o disciplină bine conturată, în cadrul căreia se profilează domenii de cercetare distincte, iar din economistul care realizează modele, un specialist cu un profil bine conturat. Econometricianul este specialistul care reuşeşte să asimileze şi să integreze elemente din teoria economică, matematică şi statistică, pentru a face o punte de legătură între fenomenele reale din practică şi necesitatea de a cunoaşte, de a explica modul de manifestare a acestor fenomene.

În contextul actual, specialistul în econometrie nu mai poate avea pretenţia că poate acoperi toate direcţiile de cercetare deschise recent şi amplificate în permanenţă. Manualele şi cărţile de econometrie se împart deja în lucrări de tip *basic econometrics* (introduceri în econometrie sau lucrări de bazele econometriei) şi *advanced econometrics* (econometrie avansată). De asemenea, în lucrările recente se disting diverse direcţii precum: econometria seriilor de timp, econometria spaţială, econometria datelor calitative, econometria datelor de tip panel, dacă avem drept criteriu tipul variabilelor pe care se bazează modelarea econometrică. După tipul sau natura fenomenului studiat, întâlnim lucrări de specialitate de tipul: econometrie financiară, bancară şi bursieră, macro-econometrie şi micro-econometrie.

Lucrarea de faţă se înscrie în categoria *basic econometrics* şi se doreşte a fi un ghid eficient la dispoziţia specialiştilor din practică, studenţilor şi profesorilor care vor să-şi însuşească noţiunile fundamentale ale teoriei şi metodologiei modelării fenomenelor economice şi sociale.

Cartea este rezultatul unui număr consistent de ani de muncă de cercetare în domeniu şi de activitate didactică de predare, susţinere de laboratoare şi seminarii la disciplinele Econometrie şi Econometrie financiară, în cadrul Facultăţii de Economie şi Administrarea Afacerilor, Universitatea „Al.I. Cuza" din Iaşi.

Structura lucrării este una clasică. În debut, sunt prezentate câteva probleme teoretice şi metodologice specifice disciplinei. La finalul acestui capitolul este prezentată o succintă introducere în R. Urmează un capitol în care sunt prezentate câteva noţiuni de teoria probabilităţilor şi statistică matematică necesare modelării econometrice. Celelalte capitole tratează sistematic problematica specifică modelării econometrice pe tipuri de modele: liniar simplu, liniar multiplu, neliniare, cu variabile alternative, problematica testării ipotezelor modelului clasic de regresie şi modele pentru serii timp. În fiecare capitol, sunt prezentate exemple cu date statistice reale utilizând pachetul program R.

Iaşi, iulie 2017 *Dănuţ-Vasile Jemna*

Capitolul 1

ELEMENTE TEORETICE ŞI METODOLOGICE

Una dintre cele mai importante inovaţii care au avut loc în domeniul economic în secolul al XX-lea a fost, cu siguranţă, econometria. Folosind puterea cifrelor, cu ajutorul statisticii şi matematicii, noii specialişti au schimbat modul de a gândi şi de a înţelege realitatea economică. Plusul de cunoaştere oferit de evaluarea empirică a teoriilor economice, succesul cu care realitatea economică era explicată cu ajutorul modelelor matematice şi statistice aveau să cucerească pe mulţi, să producă polemici şi dispute între economişti. O artă şi o ştiinţă a utilizării datelor şi teoriilor economice se dezvoltă cu rapiditate şi îşi câştigă un loc special atât în spaţiul academic, cât şi în rândul practicienilor.

Econometria este o disciplină relativ nouă ce se constituie ca urmare a eforturilor economiştilor de a explica legăturile dintre fenomenele economice şi dinamica acestora în timp, utilizând date statistice şi instrumentarul ecuaţiilor matematice. Munca cercetărilor s-a concretizat mai întâi ca parte a lucrărilor de economie politică, însă treptat specialiştii au fost preocupaţi să dezvolte şi să rafineze un instrument metodologic coerent. Astfel, s-a conturat o disciplină particulară cu un specific aparte.

Obiectivul acestui capitol este acela de a contura câteva elemente introductive ale econometriei ca disciplină distinctă. Mai precis, prezentăm succint câteva idei despre natura şi problemele cercetării econometrice.

Capitolul este deschis cu o serie de exemple şi provocări pe care le ridică econometria pentru teoria şi practica economică. De asemenea, în acest capitol, sunt abordate o serie de elemente teoretice cu privire la conceptul de model econometric. La baza acestei analize se află teoria privitoare la conceptul de *model* în general, care este aplicată apoi pentru realitatea economică (se obţine astfel

conceptul de *model economic*). În finalul capitolului, sunt prezentate o serie de probleme metodologice ale cercetării econometrice.

1.1. Provocări pentru cercetare

Pentru cititorul care nu este de specialitate, econometria poate reprezenta o disciplină teoretică, abstractă și dificil de abordat. Câteva exemple concrete unde econometria oferă rezultate și soluții cu mare putere de cunoaștere și aplicare practică pot înlesni și provoca cititorul la studiul acestei discipline.

În scopul unei evaluări intuitive a puterii cognitive a econometriei, în debutul capitolului, ne propunem o analiză a câtorva situații practice unde econometria aduce rezultate concrete și de mare importanță pentru viața economică și socială.

Exemplul 1

Să presupunem că asistăm la schimbarea managerului unei firme care are un număr mare de angajați (de exemplu, peste 100 de persoane). Pentru noul conducător al firmei, econometria oferă posibilitatea de a răspunde la o serie de întrebări de genul: care sunt criteriile după care managerul anterior a efectuat plata salariilor? Care este contribuția fiecărui factor la variația efectivului salariului de la un angajat la altul și care este ordinea de importanță a acestor factori? Există discriminare de gen sau după un alt criteriu la plata salariilor? Cum răspunde rezultatul economic la variația factorilor săi determinanți? Care este rezultatul publicității asupra vânzărilor?

Analiza relațiilor dintre un efect sau rezultat economic și factorii care îl determină reprezintă baza cercetării econometrice. Figura 1.1 ne permite vizualizarea unei asemenea relații între salariul anual ($) și o cauză sau un factor care îl determină, așa cum este educația (ani de studiu). Ceea ce se poate observa ușor este faptul că o creștere a nivelului de educație determină o creștere a mărimii salariului, adică se poate spune că ambele variabile au același sens de variație. De asemenea, se poate observa o schimbare a nivelului rezultatului atunci când se trece de un anumit prag al factorului de influență. Pentru un nivel redus de educație (8-12 ani), creșterea salariului se realizează cu o viteză relativ mică, în timp ce, pentru un nivel ridicat al educației (16-20 ani), salariul crește mai repede.

Econometria permite identificarea „modelului" sau „logicii" care explică variația salariului în funcție de variația nivelului de educație. În mod concret, se poate aprecia viteza de variație a salariului, adică cu cât variază absolut nivelul salariului la o variație cu o unitate a educației. De asemenea, se poate estima elasticitatea salariului în raport cu educația (variația relativă a salariului la o variație cu un procent a educației).

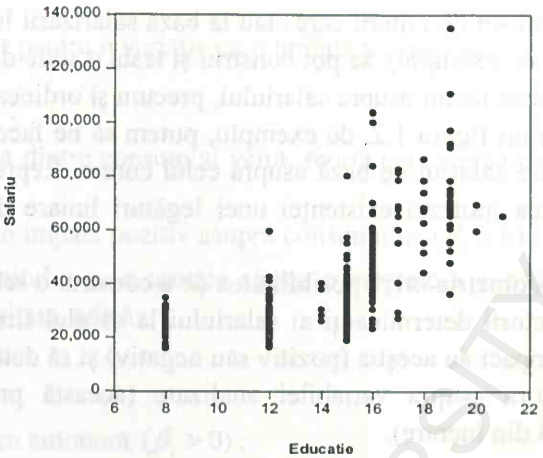


Figura 1.1. *Repartiţia angajaţilor după Salariu ($/an) şi Educaţie (ani de studiu)*
Sursa: Realizată de autor (date convenţionale)

Un asemenea rezultat este important pentru manager din mai multe puncte de vedere. Pe de o parte, econometria oferă posibilitatea construirii unei explicaţii consistente cu privire la relaţia dintre cele două variabile. O asemenea schemă explicativă poate conduce la decizii concrete, precum păstrarea sistemului de salarizare sau modificarea acestuia. Pe de altă parte, evaluarea relaţiei dintre cele două variabile permite realizarea unor predicţii în cazul unor noi angajaţi. Mai precis, econometria permite estimarea unui salariu pentru un nou angajat, având la dispoziţie „modelul" după care se plătesc angajaţii în firmă.

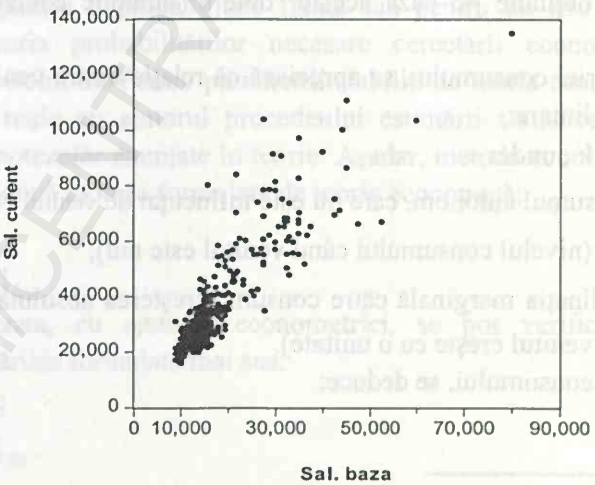


Figura 1.2. *Repartiţia angajaţilor după Salariul de bază ($) şi Salariul curent ($)*
Sursa: Realizată de autor (date convenţionale)

BCU IASI CENTRALA UNIVERSITY LIBRARY

Cu ajutorul unui set de criterii care stau la baza salarizării în firmă (așa cum este salariul de bază, de exemplu), se pot construi și testa o serie de ipoteze privind rolul și impactul fiecărui factor asupra salariului, precum și ordinea de importanță a acestora. Dacă analizăm figura 1.2, de exemplu, putem să ne facem o idee despre impactul pe care îl are salariul de bază asupra celui curent. Reprezentarea grafică ne permite formularea ipotezei existenței unei legături liniare directe între cele două variabile.

În sinteză, econometria oferă posibilitatea de a construi o schemă explicativă care să identifice factorii determinanți ai salariului la nivelul firmei analizate, să identifice ce tip de impact au aceștia (pozitiv sau negativ) și să determine o ierarhie a importanței acestora asupra variabilei analizate (această problematică face obiectul capitolului 4 din lucrare).

Exemplul 2

Pentru un student la facultatea de economie, o mare provocare este aceea de a analiza dacă teoriile formulate de marii economiști chiar funcționează în realitate și dacă anumiți parametri economici ating valorile preconizate de specialiști. Putem să ne întrebăm, de exemplu, dacă celebra teorie a consumului formulată de J.M. Keynes[1] este valabilă în condițiile socio-economice dintr-un spațiu geografic dat și pentru o perioadă istorică concretă. În acest caz, scopul cercetării este acela de a testa validitatea empirică a unei teorii. Econometria permite atingerea acestui scop printr-un demers care presupune utilizarea datelor statistice (a măsurătorilor efectuate pentru fenomenele economice considerate) și evaluarea concordanței dintre rezultatele obținute pe baza acestor date și anumite condiții stabilite de teorie.

În cazul teoriei consumului, se apreciază că relația dintre venit și consum ia forma unei ecuații liniare:

$C = \beta_0 + \beta_1 \cdot V$, unde:

- β_0 este consumul autonom, care nu este influențat de venitul curent;
- $\beta_0 = C\big|_{V=0}$ (nivelul consumului când venitul este nul);
- β_1 este înclinația marginală către consum (creșterea absolută a consumului atunci când venitul crește cu o unitate).

Din ecuația consumului, se deduce:

[1] Teoria este formulată în celebra sa lucrare *The General Theory of Employment, Interest and Money*, Palgrave Macmillan, 1936. În esență, teoria afirmă că „as income increases, consumption increases but not by as much as the increase in income" (o creștere a venitului curent atrage după sine o creștere a consumului, dar într-o măsură mai mică decât a celui dintâi).

$\beta_1 = \frac{dC}{dV}$, iar pentru o variaţie cu o unitate a venitului, rezultă:

$\beta_1 = dC\big|_{dV=1}$.

Pentru relaţia dintre consum şi venit, teoria economică propune următoarele ipoteze:

- venitul are un impact pozitiv asupra consumului ($\beta_1 > 0$);
- creşterea venitului cu o unitate atrage o creştere a consumului, dar cu mai puţin de o unitate, adică:
  $dC < dV$ sau
  $\beta_1 < 1$;
- există consum autonom ($\beta_1 > 0$);
- consumul creşte proporţional cu venitul (liniaritatea relaţiei dintre cele două variabile).

Având aceste condiţii, studentul la facultatea de economie are posibilitatea să testeze teoria cu ajutorul metodei modelării econometrice. Mai precis, se poate răspunde la întrebări de genul: care este nivelul înclinaţiei marginale către consum în România, în anul 2017? Este acesta mai mare decât cea din Bulgaria? Este mai mare de 70%? Demersul ce permite răspunsul la asemenea întrebări are la bază utilizarea datelor statistice reale care privesc o colectivitate ce respectă legile statisticii (în special, Legea numerelor mari), precum şi utilizarea unui instrument specific numit model econometric. Testarea ipotezelor se realizează în termeni statistici, adică pentru o probabilitate fixată sau pentru un risc asumat (câteva elemente de teoria probabilităţilor necesare cercetării econometrice vor fi prezentate în capitolul 2). Pentru parametrii stabiliţi de teoria economică, se obţin valori statistice reale cu ajutorul procedeului estimării statistice. Aceste valori permit testarea ipotezelor enunţate în teorie. Aşadar, metoda econometrică permite verificarea celor două ipoteze formulate de teoria economică:

- $\beta_0 > 0$;
- $0 < \beta_1 < 1$.

De asemenea, cu ajutorul econometriei, se pot verifica ipoteze care corespund întrebărilor formulate mai sus:

- $\beta_1 > 0{,}7$;
- $\beta_{1Ro} > \beta_{1Bg}$.

BCU ACADEMIA DE STUDII ECONOMICE LIBRARY

1.2. Ce este econometria?

Definițiile date pentru econometrie comportă o serie de accente de la un autor la altul din cel puțin două motive. Primul este legat de evoluția conceptuală și de schimbările de viziune care au avut loc de-a lungul timpului, iar al doilea motiv este legat de formația specialistului care a propus o definiție pentru econometrie. Așa cum observă P. Kennedy, econometricienii joacă mai multe roluri[2] care aparțin unor domenii de specialitate bine definite: ei sunt teoreticieni economiști, matematicieni, statisticieni, specialiști în conturile naționale și contabilitate, iar în definițiile date de ei econometriei apar unele accente din aceste domenii. O definiție de lucru însă se poate realiza pornind de la etimologia termenului și de la viziunea propusă de părinții acestei discipline în prima parte a secolului al XX-lea.

Definiție

Conform etimologiei termenului, *econometrie* înseamnă măsurarea fenomenelor economice sau abordarea cantitativă a realității economice. Termenul a fost consacrat în limba engleză (*econometrics*), având la bază termenii *measurement in economics* sau *economic metrics*.

Potrivit fondatorului Societății de Econometrie, R. Frisch, termenul[3] *econometrie* apare în literatura de specialitate, la începutul secolului al XX-lea, într-o lucrare germană mai puțin cunoscută. La apariția sa, în 1910, termenul *econometrie* însemna descrierea datelor economice cu ajutorul matematicii. Scopul descriptiv era atins prin prezentarea grafică, geometrică a datelor.

R. Frisch împreună cu fondatorii Societății de Econometrie (1930) au definit econometria într-o perspectivă mai complexă, adăugând sensului originar al termenului dimensiunea explicativă și pe cea predictivă. Părinții disciplinei au dat econometriei următoarea definiție: analiza cantitativă a fenomenelor economice, pe baza teoriei economice și a datelor de observație, utilizând metode specifice ale inferenței statistice („the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference"[4]).

Econometria este o disciplină care s-a dezvoltat prin integrarea unor elemente specifice din mai multe științe: economie politică (teoriile și mecanismele economice), matematică economică (modelele matematice), statistică matematică și statistică economică (instrumentele de culegere și prelucrare a datelor și metodele de inferență).

---

[2] P. Kennedy, *A Guide to Econometrics*, MIT Press, 1998, p. 1.

[3] R. Frisch, *Econometrica*, vol. 4 (1), 1936, p. 95.

[4] P. Samuelson, T. Koopmans, R. Stone, „Report of the Evaluative Committee for Econometrica", *Econometrica*, 1954, p. 142.

Scopul econometriei este crearea suportului empiric pentru formularea şi verificarea teoriilor economice[5], precum şi pentru elaborarea deciziilor de politică economică. Acesta este realizat prin atingerea următoarelor obiective:

- descrierea şi explicarea dependenţelor dintre fenomenele economice;
- explicarea variaţiei în timp a fenomenelor economice;
- explorarea realităţii economice cu ajutorul datelor statistice pentru a propune noi teorii şi ipoteze;
- testarea ipotezelor elaborate în teoria economică;
- predicţia fenomenelor economice.

Complexul de elemente din economie, matematică şi statistică sunt convenabil sintetizate în noua disciplină, în acord cu obiectivele enunţate mai sus, astfel că econometria depăşeşte stadiul unei simple colecţii de elemente din alte ştiinţe. Economia politică admite, în principal, o abordare calitativă asupra vieţii economice. Această perspectivă este completată de econometrie cu o abordare de tip cantitativ. Matematica economică oferă modele matematice, teoretice, pe care econometria le transformă în modele econometrice, adică modele care pot fi estimate şi testate cu ajutorul statisticii matematice, pe baza datelor empirice oferite de statistică.

Econometria este o disciplină metodologică care s-a dezvoltat îndeosebi pe baza realizărilor din cercetarea statistică cu privire la: estimarea parametrilor modelelor legăturilor dintre fenomenele economice, testarea ipotezelor statistice cu privire la teoriile economice, analiza şi prognoza în timp a fenomenelor economice, fundamentarea politicilor de decizie economică.

Componente

În funcţie de tipul datelor statistice[6] pe care le foloseşte această disciplină, putem identifica două componente ale econometriei: econometria seriilor de timp *(time-series econometrics)* şi econometria datelor din anchete *(cross-sectional econometrics)*. Când cele două metode de analiză sunt combinate, se obţine analiza de tip panel, care se divide în alte două tipuri de analiză: analiza pentru acelaşi eşantion la diferite momente de timp *(panel analysis)* şi analiza pentru eşantioane diferite *(pooled cross-section data)*.

Coordonate istorice

Ca disciplină ştiinţifică, econometria are o istorie scurtă, ce a debutat cu înfiinţarea *Societăţii de Econometrie*, în anul 1930. Rădăcinile disciplinei sunt mai

---

[5] G.S. Maddala, *Introduction to Econometrics*, John Wiley & Sons, 2001, p. 3.
[6] D. Gujarati, *Basic Econometrics*, McGraw Hill, 2004, pp. 25-28.

adânci şi sunt legate de apariţia calculului probabilistic şi de încercările specialiştilor de a explica fenomene de masă pe baza inferenţei statistice.

Principalele momente din istoria ştiinţei care au pregătit terenul pentru apariţia econometriei sunt: crearea *Şcolii Aritmeticii Politice*, în Anglia secolului al XVII-lea, înfiinţarea laboratoarelor biometrice engleze la sfârşitul secolului al XIX-lea şi naşterea *Societăţii Econometrice*, în anul 1930. Părintele Şcolii Aritmeticii Politice engleze, W. Petty, este un prim reper pentru înţelegerea econometriei. De numele lui se leagă studii academice care îmbină teoria cu datele de observaţie, cu scopul de a explica anumite fenomene şi legături dintre fenomene sau de a cunoaşte fenomene de masă, pe baza unor observaţii parţiale.

Econometria datorează poate cel mai mult părinţilor englezi ai şcolii de biometrie şi de statistică: F. Galton, K. Pearson, R.A. Fisher, F.Y. Edgeworth. Principalele concepte şi metode utilizate în analiza legăturilor dintre fenomene au fost lansate de aceşti gânditori iluştri. Acestora le datorăm conceptele de *regresie* şi *corelaţie*.

Sub aspect instituţional, pentru econometrie este important anul 1930, an în care se înfiinţează *Econometric Society* la Cleveland, în Ohio, USA. Iniţiativa Societăţii de Econometrie îi aparţine lui Irving Fisher, primul preşedinte al societăţii[7], şi economistului norvegian Ragnar Frisch, cel care a introdus termenul *econometrie (econometrics)*, în anul 1926. Ca membri ai societăţii, menţionăm câteva personalităţi de seamă: Ragnar Frisch, Irving Fisher, R.A. Fisher, Jan Timbergen, Trygve Haavelmo.

Cu ocazia unei evaluări a activităţii Societăţii de Econometrie, în 1954, în raportul comitetului de conducere, este prezentată una dintre multele definiţii ale econometriei preluate de specialişti de-a lungul timpului:

Econometria, rezultatul unor anumite concepţii asupra rolului economiei, constă în aplicarea statisticii matematice asupra datelor economice cu scopul de a oferi suportul empiric pentru modelele construite de matematica economică şi pentru a obţine rezultate cantitative.[8]

În anul 1933, Societatea de Econometrie editează prima revistă de econometrie cu numele *Econometrica*, avându-l ca redactor-şef pe R. Frisch.

---

[7] www.econometricsociety.org

[8] P.A. Samuelson, T.C. Koopmans, J.R.N. Stone, „Report of the Evaluative Committee for Econometrica", *Econometrica*, vol. 22 (2), April 1954, pp. 141-146 („**Econometrics**, the result of a certain outlook on the role of economics, consists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results").

Dezvoltări ulterioare

Atât Societatea de Econometrie, cât şi revista *Econometrica* au rezistat până astăzi prin munca şi contribuţia unor oameni de ştiinţă importanţi: C.W. Cobb şi P.H. Douglas (în domeniul producţiei), K. Schultz, P.A. Samuelson (în domeniul cererii), J. Timbergen, T. Haavelmo, R. Frisch, L.R. Klein, H. Theil, J.M. Keynes (în domeniul modelelor macroeconomice).

Econometria a devenit o disciplină academică importantă, astfel că în curriculum-ul oricărei facultăţi de economie se regăsesc unul sau mai multe cursuri care prezintă elemente fundamentale de econometrie (*Basic Econometrics*), elemente avansate (*Advanced Econometrics*) sau cursuri mai specializate (*Financial Econometrics, Panel Data Econometrics*).

1.3. Probleme teoretice

În această secţiune, sunt vizate principalele concepte care stau la baza econometriei ca disciplină de specialitate. Ne referim în principal la conceptul de model econometric şi la componentele sale (variabile şi parametri), precum şi la instrumentarul care permite construirea unor modele pe baza datelor reale (variabile, estimatori, ipoteze, teste).

1.3.1. Definirea conceptului de model econometric

În econometrie, conceptul de *model* a fost preluat din teoria economică şi din matematică. Pentru economia politică, modelul economic este o schemă, un mecanism care explică modul în care funcţionează economia ca întreg sau un sector al economiei. Modelul matematic este un sistem formal, determinat de o ecuaţie sau de un sistem de ecuaţii cu proprietăţi bine definite.

a. Conceptul de model în economie

Termenul *model* are o semnificaţie primară care vizează o realitate fizică. Acesta trimite la un obiect material sau la o reprezentare a unei structuri la o anumită scară. În economie, termenul îşi pierde substanţa fizică, dar îşi păstrează puterea de reprezentare a realităţii ca schemă explicativă. Odată construite, modelele pot fi tratate ca şi quasi-obiecte empirice.

Pe linia tradiţiei cercetărilor din fizică, a înţelege un fenomen înseamnă a construi un model care să imite acel fenomen. Dintr-un anumit punct de vedere, modelele economice au funcţionat ca nişte mecanisme, după exemplele din fizică. A alege o formă matematică adecvată înseamnă a alege un număr de parametri, de valori, astfel încât mecanismul să reproducă anumite fapte ale fenomenului. Valorile parametrilor sunt obţinute pe baza datelor de observaţie, în general, prin procedeul de estimare.

Practica economică este dominată de construirea de modele. Pentru a evalua realitățile și politicile economice, se construiesc și se utilizează seturi de date, se realizează evaluări cantitative. Fenomenele economice nu sunt direct observabile, ci sunt analizate cu ajutorul datelor statistice. De asemenea, teoriile economice nu vizează observații particulare, ci sunt formulări despre fapte și fenomene privite ca ansambluri de fapte și fenomene individuale. Acestea din urmă nu sunt direct observabile, ci sunt analizate pe baza unor măsurători sau date de observație.

Modelele economice pot fi privite ca niște instrumente de observare și de măsurare a realității. Modelele nu sunt, pur și simplu, derivate din teoria economică, ci sunt construite prin integrarea ideilor teoretice și a datelor empirice despre un fenomen economic. Deoarece teoriile sunt incomplete și au caracter conjectural (ipotetic), fundamentarea lor se realizează pe baza datelor de observație din realitatea economică.

b. Modelul econometric. Descriptiv și explicativ în analiza fenomenelor social-economice

După posibilitățile oferite de cercetarea statistică, analiza cantitativă a fenomenelor sociale și economice se poate realiza la două niveluri, care vizează două posibilități de cunoaștere: descriptivă și explicativă.

Descrierea fenomenelor sociale și economice se realizează cu ajutorul indicatorilor statistici de bază, determinați pe serii de date observate din realitate. Menționăm aici indicatorii tendinței centrale, indicatorii dispersiei, indicatorii formei și concentrării, indicatorii factoriali ai dispersiei etc. Rezultatele obținute prin analiza descriptivă oferă, de obicei, pe lângă informațiile utile pentru cunoașterea realității studiate, datele și informațiile necesare explicării acelei realități.

Obiectivul explicativ al cercetării cantitative a fenomenelor sociale și economice se realizează cu ajutorul modelelor.

Termenul *model* are mai multe conotații:

- descriptivă, ce implică ideea de schemă care descrie realitatea în anumiți parametri;
- cauzală sau explicativă, care sugerează ideea unui lanț de explicații;
- normativă sau prescriptivă, care conturează ideea unui demers sau a unui traseu ce trebuie urmat;
- probabilistă, care trimite la un sistem ipotetic de distribuire a variabilelor aleatoare.

În teoria și practica economică, termenul *model* este utilizat cu cea de-a doua conotație, explicativă. *Modelul* este o prezentare formalizată, schematică a realității economice studiate, cu scopul de a explica modul în care aceasta se manifestă.

Modelul econometric ia forma unei ecuaţii (sistem de ecuaţii) cu două sau mai multe variabile statistice. În econometrie, *modelul* reprezintă instrumentul prin care se încearcă să se explice realitatea studiată în dimensiunile sale fundamentale. Obiectivul construirii acestor modele este de a înţelege şi de a explica realitatea economică, în vederea luării unor decizii practice concrete.

Dacă este să ne referim la un exemplu clasic, modelul consumului al lui Keynes reprezintă o funcţie care explică în ce mod creşterea veniturilor populaţiei determină o creştere a consumului. Modelul are la bază ecuaţia matematică:

$Y = \beta_0 + \beta_1 X$, unde:

- Y reprezintă consumul;
- X este venitul;
- β_0 este consumul autonom;
- β_1 este înclinaţia marginală către consum.

Acest model reprezintă o schemă simplificată a realităţii economice privind consumul. Modelul presupune o dependenţă liniară sau proporţională a consumului de factorul cauzal venit. În mod cert, consumul nu este determinat doar de venit, ci de un număr mare de factori, aceştia putând fi luaţi în considerare, explicit sau nu, în model. În cazul teoriei lui Keynes însă, modelul presupune un singur factor explicativ, venitul, care este considerat determinant şi suficient.

Ca schemă simplificată a realităţii, modelul este construit cu obiective clare, iar alegerea factorilor explicativi în model depinde de aceste obiective ale cercetării. De asemenea, numărul factorilor din model depinde de posibilităţile de măsurare a realităţii economice.

Econometria este o ştiinţă empirică analogă astronomiei. Sarcina specialistului în econometrie nu este să dovedească că o teorie este adevărată sau nu, ci să caute modele satisfăcătoare pentru acea teorie. Modelele econometrice operează atât ca scheme explicative ale realităţii, cât şi ca instrumente de investigare a acesteia.

1.3.2. Noţiuni, termeni şi notaţii

În modelarea econometrică, pe lângă conceptul de *model*, apar o serie de noţiuni şi termeni specifici care provin din statistică şi matematică: variabile, parametri, ipoteze statistice, teste statistice etc.

1. Variabile

Discuţia privind variabilele utilizate în modelarea econometrică admite cel puţin două componente. Mai întâi, este vorba despre o distincţie între tipuri de variabile ce corespund celor trei surse care compun econometria. Mai apoi, este

vorba despre o tipologie a variabilelor care este strâns legată de instrumentul metodologic, și anume modelul econometric.

În acord cu definiția econometriei ca un mix de discipline (economie, matematică și statistică), este important să distingem între trei tipuri de variabile cu care lucrează cercetătorul ce face apel la metodologia econometrică. Prezentăm aceste tipuri de variabile în tabelul 1.1.

Tabelul 1.1. *Tipuri de variabile utilizate în econometrie*

	Economice	Teoretice	Statistice
Caracteristici	Stau la baza teoriilor și modelelor economice. Au un grad mare de abstractizare.	Sunt funcții matematice cu proprietăți bine definite.	Sunt instrumente care măsoară realitatea economică. Pot fi numerice și nominale.
Exemple	Venit, consum, dobândă, salariu, șomaj, inflație.	Variabile aleatoare (între care se află și estimatorii), funcții continue și discrete.	Rata șomajului (%), salariul mediu net (lei), PIB pe locuitor (lei).

Sursa: Realizat de autor

Cercetarea cu ajutorul econometriei presupune identificarea variabilelor economice de interes. Acestea sunt definite în cadrul teoriilor economice, deseori având un caracter abstract. Pentru modelarea econometrică, este necesar ca pentru fiecare variabilă economică de interes să se identifice una sau mai multe variabile statistice. Acestea se referă la populații reale, finite și reprezintă instrumentele cu care este măsurată realitatea economică. Acest demers este deseori un procedeu dificil pentru că este legat de posibilitățile de măsurare și de disponibilitatea datelor statistice (statistica oficială, anchetele statistice, sistemul financiar-contabil etc.). Variabilele statistice trebuie identificate în funcție de tipul lor (după natura lor și modul de exprimare) și după sursa datelor disponibile. Pentru diverse probleme teoretice și metodologice, pe lângă variabilele statistice și economice, sunt utilizate și variabile teoretice, construite pe populații ipotetice. În această categorie intră variabilele aleatoare ce sunt utilizate în inferența statistică.

La nivelul unui model econometric, se pot identifica mai multe tipuri de variabile. Vom exemplifica această tipologie cu ajutorul modelului econometric ce stă la baza teoriei consumului. Modelul econometric propus de Keynes este dat prin ecuația:

$$Y = \beta_0 + \beta_1 X + \varepsilon .$$

În acest model econometric, apar următoarele tipuri de variabile:

- *variabila dependentă.* Este denumită și *variabilă explicată*, *endogenă*, *răspuns*, *rezultativă* și cuantifică un fenomen economic complex, determinat

de o serie de factori (în acest caz, variabila dependentă este consumul). Această variabilă se notează, de obicei, cu litera *Y*.

- *variabila independentă* sau *variabila explicativă, exogenă, stimul, predictor* sau *variabilă de control*. Se notează cu litera *X* sau cu X_j, $j = \overline{1,p}$, dacă în model sunt *p* variabile independente. Variabila independentă este o variabilă statistică prin care se măsoară acţiunea unui factor economic asupra rezultantei (în modelul analizat, variabila independentă este venitul). În poziţia de variabilă independentă, în anumite modele econometrice, poate apărea variabila timp, care dă caracterul dinamic al modelului.
- *variabila aleatoare* sau *termenul aleator*. Este numită şi *variabilă reziduală* sau *eroare* şi sintetizează influenţa tuturor factorilor care nu apar explicit în model, precum şi erorile de modelare, determinate de procesul de culegere a datelor şi de metoda statistică. Aceasta este o variabilă aleatoare şi se notează cu ε.

Observaţie

În modelul econometric, variabila dependentă este considerată o variabilă aleatoare (pentru că depinde de variabila reziduală, care este o variabilă aleatoare).

2. Parametri. Estimatori. Estimaţii

Parametrii modelului econometric se mai numesc şi *coeficienţi de regresie* şi reprezintă acele mărimi fixe, dar necunoscute, care apar în modelul econometric, pe lângă variabile, în diverse expresii.

Se notează cu litere greceşti, β_j, pentru $j = \overline{0,p}$, şi constituie obiectul procesului de estimare. Într-un model, de regulă, există $k = p+1$ parametri, unde *p* este numărul de variabile independente.

Estimatorii sunt variabile aleatoare construite pe spaţiul de selecţie, cu scopul de a estima parametrii modelului. Pentru estimatori, este important să se cunoască legea de repartiţie şi proprietăţile acestora. Estimatorii se notează cu $\hat{\beta}_j$, $j = \overline{0,p}$.

Estimaţiile sunt valori posibile ale estimatorilor pe spaţiul de selecţie. La nivelul unui eşantion sau set de date statistice, estimaţia este o valoare cunoscută, calculată pe baza datelor de observaţie. Estimaţiile se notează cu litere latine, b_j, $j = \overline{0,p}$.

3. Ipotezele statistice

În modelarea econometrică, sunt formulate o serie de ipoteze care provin din teoria economică sau din obiectivele cercetării. Toate acestea iau forma unor

ipoteze statistice, adică a unor presupuneri cu privire la legea de distribuție a unei variabile sau cu privire la parametrii modelului ori alți diverși parametri, cum ar fi indicatorii de corelație.

De asemenea, în econometrie, este consacrat un set de ipoteze cu privire la variabilele care compun modelul econometric. Aceste ipoteze sunt presupuneri cu privire la legea de repartiție a variabilelor aleatoare din modelul econometric, precum și cu privire la relația dintre variabilele independente care apar în model. Aceste presupuneri se numesc *ipotezele modelului clasic de regresie* și fac obiectul unei etape metodologice distincte (subiect care este tratat în capitolul 7 al acestei cărți).

4. Testele statistice

Testele statistice utilizate în econometrie sunt procedee la finalul cărora, pe baza unei reguli, se ia decizia de a respinge sau de a nu respinge ipoteza supusă testării. La baza testelor se află *statisticile*, adică variabile aleatoare cu legi de repartiție cunoscute și complet specificate. În baza proprietăților acestor variabile aleatoare, se ia o decizie care are caracter probabilistic. În capitolul al doilea al acestei lucrări, sunt prezentate cele mai importante *statistici* utilizate în modelarea econometrică.

1.4. Probleme metodologice

Pentru cercetarea econometrică, este important să se cunoască etapele demersului modelării și să se clarifice problemele legate de tipul și calitatea datelor statistice utilizate în modele. Etapele demersului metodologic sunt cele specifice unei cercetări științifice de tip cantitativ, iar datele utilizate provin din surse autorizate.

1.4.1. Demersul metodologic al cercetării econometrice

În sinteză, principalele etape ale cercetării econometrice sunt prezentate mai jos:

- *formularea unei probleme sau a unui set de ipoteze.* Această etapă se realizează pe baza teoriilor și a cercetărilor anterioare, dar și pe baza datelor de observare culese pentru un anumit fenomen.
- *formalizarea problemei într-un model.* În această etapă, se propune un model economic sau matematic (funcțional) pentru teoria sau ipotezele propuse și apoi se specifică un model econometric (model cu variabile statistice observabile).
- *obținerea datelor pentru modelare.* Este o etapă deosebit de importantă de care depinde calitatea rezultatelor. În această etapă, se specifică tipul datelor

şi metodele prin care acestea pot fi obţinute, care sunt apoi culese după procedeele alese.

- *estimarea parametrilor modelului econometric.* Estimarea presupune, mai întâi, alegerea metodelor de estimare şi a estimatorilor, iar apoi, se trece la aplicarea acestora pe setul de date disponibile.
- *testarea ipotezelor.* Vizează atât testarea parametrilor modelului, cât şi a modelului în sine, precum şi a condiţiilor şi proprietăţilor cerute de teoria economică şi de metodologia statistică.
- *predicţia fenomenului.* Reprezintă proiectarea fenomenului analizat pentru o perioadă de timp viitoare sau proiectarea unor scenarii pentru date specificate ale variabilelor factoriale, având la bază modelul estimat.
- *utilizarea modelului în practica economică.* Vizează luarea unor decizii de politică economică, realizarea controlului unor activităţi etc.

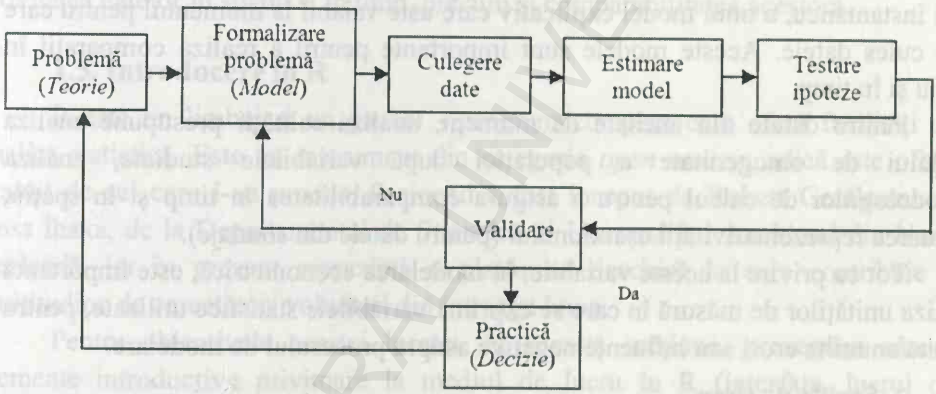


Figura 1.3. *Demersul metodologic al modelării econometrice*
Sursa: Realizată de autor

Schematic, demersul cunoaşterii econometrice poate fi reprezentat schematic aşa cum se prezintă Figura 1.3. Metodologia ne arată că, în econometrie, este esenţială corelarea rezultatelor modelării cu practica economică. Uneori, rezultatele modelării înseamnă revenirea asupra ipotezelor sau teoriei economice, descoperirea de noi teorii sau reformularea ipotezelor de plecare. Alteori, rezultatul cercetării este utilizat în practică pentru formularea de strategii, pentru luarea de decizii. De asemenea, modelul obţinut este supus procesului de validare, ceea ce poate însemna deseori revenirea asupra etapei de elaborare a modelului, de identificare a variabilelor şi a formei legăturii dintre variabile.

BCU - CENTRALA UNIVERSITARĂ LIBRARY

1.4.2. Tipuri de date statistice utilizate în econometrie

În econometrie, se utilizează, de regulă, trei tipuri de date: datele furnizate de anchetele statistice la un moment dat, seriile de timp şi datele obţinute prin anchete de tip panel, care combină seriile de timp cu datele observate la un moment dat în anchete. Pentru fiecare tip de date, calitatea reprezintă o problemă fundamentală. Calitatea datelor se stabileşte în funcţie de tipul datelor, de sursele acestora, de metodele de culegere, de completitudinea şi accesibilitatea lor.

1. Seriile de moment

Aceste serii de date provin din diferite tipuri de anchete. De regulă, datele pentru modelare se obţin prin cercetări pe bază de sondaj statistic, dar există şi date disponibile din cercetări exhaustive de tip recensământ. Datele din anchete se referă la populaţii statistice bine delimitate în spaţiu, timp şi cu privire la natura lor.

Construirea de modele cu ajutorul datelor din anchete presupune obţinerea unui instantaneu, a unui model explicativ care este valabil la momentul pentru care s-au cules datele. Aceste modele sunt importante pentru a realiza comparaţii în spaţiu şi în timp.

Pentru datele din anchete de moment, analiza calităţii presupune analiza gradului de omogenitate a populaţiei după variabilele studiate, analiza metodologiilor de calcul pentru a asigura comparabilitatea în timp şi în spaţiu, evaluarea reprezentativităţii eşantionului (pentru datele din sondaje).

Tot cu privire la aceste variabile, în modelarea econometrică, este importantă analiza unităţilor de măsură în care se exprimă variabilele statistice utilizate, pentru a evita anumite erori sau influenţe negative asupra procesului de modelare.

2. Seriile de timp

Sunt seturi de date observate pentru un fenomen la diverse momente sau intervale de timp. Aceste serii se construiesc cu ajutorul variabilelor numerice sau atributive şi sunt seturi de date sub forma unor înregistrări orare, zilnice, săptămânale, lunare sau anuale.

Calitatea seriilor de timp este analizată prin:

- evaluarea comparabilităţii datelor – în timp, metodologiile de calcul şi de observare se pot modifica;
- verificarea surselor şi a metodelor de culegere a datelor – pot exista mai multe surse de date pentru un fenomen şi anumite discordanţe între ele.

Principalele probleme ale seriilor de timp în economie sunt legate de disponibilitatea şi accesibilitatea acestora şi de gradul de agregare a datelor. Prin statistica oficială, se culeg date pentru un număr restrâns de variabile, iar în situaţii concrete de cercetare, nu este posibilă realizarea de modele, tocmai din lipsă de date statistice. De asemenea, o altă problemă este aceea că datele nu sunt

disponibile la nivel local sau regional, majoritatea datelor statistice oficiale fiind agregate la nivel naţional.

3. Seriile de tip panel

Prin anchetele de tip panel, se combină avantajele analizei de moment şi ale analizei în timp a unui fenomen. Din aceste anchete, se obţin serii de timp pentru acelaşi eşantion de unităţi sau pentru eşantioane care sunt modificate, dar care păstrează aceleaşi caracteristici ale populaţiei.

Anchetele panel permit obţinerea de serii de date de moment, dacă se consideră o anchetă realizată la un anumit moment, cât şi serii de timp, dacă se consideră rezultatele anchetelor pe o perioadă de timp. Prin anchetele panel, sunt îmbunătăţite condiţiile de calitate a datelor obţinute atât din anchete, cât şi datele de tip serii de timp. Printr-o cercetare selectivă, datele din anchetele panel permit observarea şi analiza unui fenomen în dezvoltarea sa în timp, asigurând reprezentativitatea datelor în spaţiu şi în timp, precum şi comparabilitatea acestora.

1.5. Introducere în R

R este un limbaj şi un pachet - program complex care oferă facilităţi de analiză statistică. Este un instrument din categoria *open source*, adică este oferit gratuit de cei care l-au produs. Proiectul a fost început de Robert Gentleman şi Ross Ihaka, de la Departamentul de Statistică al Universităţii Auckland din Noua Zeelandă, iar în prezent reprezintă o platformă deschisă la care contribuie o multitudine de cercetători voluntari din întreaga lume.

Pentru obiectivele acestei lucrări, în această secţiune, prezentăm câteva elemente introductive privitoare la mediul de lucru în R (interfaţa, lucrul cu comenzile, pachete, funcţii şi obiecte, tipuri de fişiere, variabile şi structuri de date etc.) şi câteva elemente privind analiza statistică în R.

1.5.1. Mediul de lucru

Programul R necesită o instalare, după ce a fost descărcat fişierul de bază de la sursa gratuită https://www.r-project.org. După instalare, la deschiderea programului apare fereastra din figura 1.4.

Comenzi şi funcţii

Toate comenzile în R se introduc în fereastra din figura 1.4 numită *Consola R*. Funcţionând ca un limbaj, este necesară cunoaşterea funcţiilor şi sintaxei pentru a da o comandă ce permite obţinerea unor rezultate dorite. O comandă poate fi una foarte simplă, ce utilizează o singură funcţie, sau una mai complexă, ce apare ca o secvenţă de funcţii.

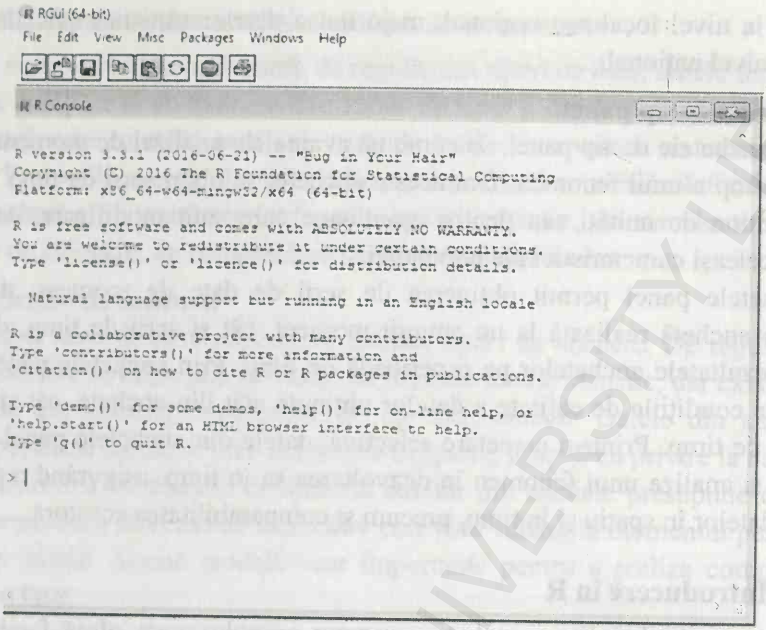


Figura 1.4. *Interfața R*
Sursa: Captură din pachetul R

Un aspect important în privința sintaxei îl reprezintă biblioteca pusă la dispoziție de R cu privire la toate funcțiile disponibile. Pentru a accesa această bibliotecă, este suficient să se scrie în consola R următorul cod simplu:

```
> ?nume funcție()
```

De exemplu, dacă se dorește să se obțină o informație despre funcția matematică *abs*, care returnează valoarea absolută a unui număr, se scrie comanda >?abs(), iar biblioteca se deschide și dă o serie de informații: o descriere a funcției („abs(x) computes the absolute value of *x*"); sintaxa (care este simplă - abs(x), unde *x* este valoarea pentru care se aplică funcția); o descriere a argumentului („numeric or complex vector or array"); alte detalii privind funcția.

Pachete (Packages)

R este structurat astfel încât, pe lângă componentele deținute la instalare, să se poată dezvolta prin adăugarea de noi pachete, care sunt disponibile gratuit pentru a fi instalate.

Asemenea pachete sunt construite de diferiți voluntari și pot fi instalate sau dezinstalate după preferințe. În meniul de bază există opțiunea *Packages* care permite atât instalarea (*Install package*), cât și încărcarea de noi pachete (*Load Package*) în spațiul de lucru.

Pentru a vedea ce pachete sunt instalate la un moment dat, se utilizează comanda: > library(). După aplicarea comenzii, se deschide o nouă fereastră în care sunt prezentate toate pachetele disponibile și informații despre acestea (a se vedea figura 1.5). Lista pachetelor disponibile se poate obține și aplicând: > search().

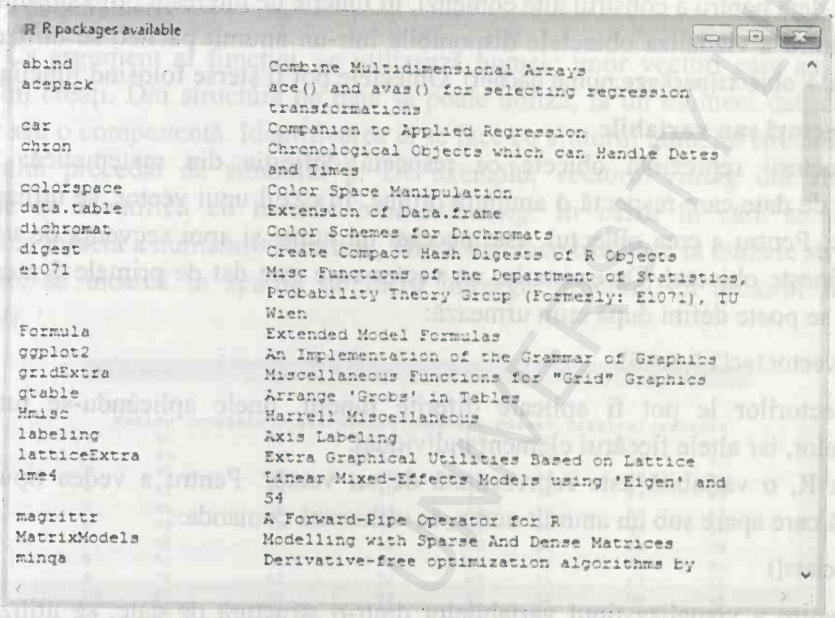


Figura 1.5. *Pachete disponibile în R*
Sursa: Captură din pachetul R

Obiecte

În R, structura operațională de bază este reprezentată de obiecte. Acestea apar ca rezultat al aplicării funcțiilor. De regulă, pentru fiecare obiect, putem atribui un nume care este memorat împreună cu informația definită în acesta (este vorba despre date sau structuri de date și rezultate ale analizelor statistice etc.). Este important de subliniat că în R se face distincție între caracterele scrise normal și cele cu majusculă (de exemplu, numele de obiecte *demo* și *Demo* diferă). Comanda de creare a obiectului presupune scrierea numelui urmat de semnul egal și o sintaxă care definește obiectul (de la o simplă valoare până la o funcție mai simplă sau mai complexă). În R, o comandă are două părți: obiectul și funcția:

> obiect<-funcție

Simbolul <- este folosit în R pentru a indica ideea de „a crea prin intermediul", de a crea obiectul prin intermediul funcției. De exemplu, dacă se tastează comanda x<-100, s-a creat un obiect cu numele *x* care memorează valoarea

100. Pentru a vedea conținutul unui obiect, se introduce în consola R numele obiectului drept comandă ce se execută tastând *Enter*.

Obiectele pot fi folosite ulterior în diverse comenzi, (de exemplu, se poate calcula radical din x utilizând comanda >sqrt(x)). R permite utilizarea unui obiect construit deja pentru a construi alte comenzi, în funcție de interesul utilizatorului.

Pentru a vizualiza obiectele disponibile într-un anumit pachet, se utilizează comanda: > objects(package:nume pachet). Obiectele pot fi șterse folosind funcția *rm*.

Vectori sau variabile

Vectorii reprezintă obiecte ce respectă definiția din matematică: sunt structuri de date care respectă o anumită ordine. În cazul unui vector, se utilizează funcția *c*. Pentru a crea obiectul, este necesar un nume și apoi secvența de valori care definește obiectul. De exemplu, un vector care este dat de primele 5 numere naturale se poate defini după cum urmează:

```
> vector1=c(1,2,3,4,5)
```

Vectorilor le pot fi aplicate diferite funcții, unele aplicându-se tuturor elementelor, iar altele fiecărui element individual.

În R, o variabilă este reprezentată de un vector. Pentru a vedea tipul de variabilă care apare sub un anumit nume, se utilizează comanda:

```
> class()
```

Pentru a vizualiza tipul variabilelor dintr-o structură de date, se utilizează comanda:

```
> sapply(dataset, class)
```

Tipurile de variabile în R sunt: *numeric* (numere reale); *integer* (numere întregi); *factor* (pot fi numerice sau categoriale de tip ordinal); *date* (de tip dată calendaristică); *character* (variabile nominale).

Funcția *sapply()* permite aplicarea aceleiași funcții menționate în argument unei întregi structuri de date, mai precis, pentru fiecare vector în parte din acea structură.

Structuri de date

Un alt obiect important în R este reprezentat de o structură de date (*data frame*). Aceasta este compusă din mai mulți vectori care au aceeași dimensiune. Structura de date seamănă cu un tabel care are mai multe coloane, iar fiecare coloană are un nume. Datele din fiecare coloană pot fi de tip numeric sau de alt tip.

În R, se pot accesa anumite seturi de date care sunt puse la dispoziție de program. Pentru a vedea numele acestor seturi de date, se utilizează comanda:

```
>data()
```

BCU CLUJ CENTRAL UNIVERSITY LIBRARY

De exemplu, pentru a vedea structura de date numită *attitude*, se tastează în consolă efectiv numele acesteia şi se pot vizualiza cele 7 coloane cu date (vezi figura 1.6).

O structură de date se construieşte cu ajutorul comenzii:

```
> data.frame()
```

Ca argument al funcţiei, se utilizează numele unor vectori care au fost în prealabil creaţi. Din structura de date se poate utiliza, la un moment dat, doar un vector sau o componentă. Identificarea ei se face cu ajutorul numelui structurii şi al vectorului precedat de simbolul $. De exemplu, vectorul *rating* din structura *attitude* se identifică cu numele *attitude$rating*. În cazul în care se doreşte utilizarea directă a numelui vectorului, fără a mai face trimitere la numele structurii de date, se încarcă în spaţiul de lucru întreaga structură cu ajutorul funcţiei *attach()*.

R R Console

```
> attitude
   rating complaints privileges learning raises critical advance
1      43         51         30       39     61       92      45
2      63         64         51       54     63       73      47
3      71         70         68       69     76       86      48
4      61         63         45       47     54       84      35
5      81         78         56       66     71       83      47
6      43         55         49       44     54       49      34
7      58         67         42       56     66       68      35
8      71         75         50       55     70       66      41
9      72         82         72       67     71       83      31
10     67         61         45       47     62       80      41
11     64         53         53       58     58       67      34
12     67         60         47       39     59       74      41
13     69         62         57       42     55       63      25
14     68         83         83       45     59       77      35
15     77         77         54       72     79       77      46
16     81         90         50       72     60       54      36
17     74         85         64       69     79       79      63
18     65         60         65       75     55       80      60
19     65         70         46       57     75       85      46
20     50         58         68       54     64       78      52
21     50         40         33       34     43       64      33
22     64         61         52       62     66       80      41
23     53         66         52       50     63       80      37
24     40         37         42       58     50       57      49
25     63         54         42       48     66       75      33
26     66         77         66       63     88       76      72
27     78         75         58       74     80       78      49
28     48         57         44       45     51       83      38
29     85         85         71       71     77       74      55
30     82         82         39       59     64       78      39
```

Figura 1.6. *Structură de date în R*
Sursa: Captură din pachetul R

Prin utilizarea parantezelor pătrate, se poate defini un set de date mai redus (sub-set de date) din structura existentă. Se poate selecta o valoare indicând linia şi coloana, o anumită linie, o anumită coloană sau un subset anume. De exemplu, prin comanda > attitude[2,3], se permite selectarea valorii ce apare în structura de date pe linia a 2-a şi coloana a 3-a. R returnează valoarea: [1] 51.

Dacă se dorește vizualizarea numelor variabilelor care apar într-o structură de date, se folosește comanda:

```
> names(dataframe)
```

În exemplul utilizat, comanda > names(attitude) returnează rezultatul:

```
[1] „rating" „complaints" „privileges" „learning" „raises" „critical" „advance"
```

O structură de date în R poate fi vizualizată separat într-o fereastră numită *data editor* utilizând funcția *fix* urmată de numele structurii de date. Pentru exemplul considerat, putem deschide fereastra editor utilizând comanda:

```
> fix(attitude)
```

Combinarea de date în structuri poate fi realizată și cu ajutorul altor funcții precum list(), cbind(), rbind(). Prima funcție creează o listă de obiecte care sunt puse într-o anumită ordine, așa cum au fost introduse la argumentul funcției. Funcțiile cbind() și rbind() au rolul de a pune împreună coloane de date sau linii de date din variabile sau structuri de date diferite.

Dacă într-o structură de date sau la nivelul unei variabile lipsesc date pentru anumite unități, în R, se realizează o codificare automată prin simbolul NA ("not available"). Acest simbol poate fi folosit în diverse funcții pentru a trata datele lipsă într-un anumit mod. De exemplu, dacă în aplicarea unei funcții se dorește ignorarea valorilor lipsă, se adaugă în sintaxa comenzii opțiunea:

```
na.rm = TRUE
```

Spațiu de lucru (*Workspace*)

Tot ce s-a lucrat în R, mai precis toate obiectele definite, se găsește în spațiul de lucru și poate fi salvat într-un fișier distinct cu ajutorul meniului aflat în linia de sus a consolei (opțiunea *File/Save Workspace*). Ca rezultat, se obține un fișier de tip *.RData.

Toate obiectele dintr-un *workspace* pot fi vizualizate cu ajutorul comenzii >objects().

Fișiere script

Toate comenzile create și utilizate într-o consolă pot fi salvate într-un fișier de tip script *.R. Pentru aceasta, se deschide o fereastră R Editor unde sunt așezate comenzile. Utilizarea lor presupune să plasăm cursorul pe linia de comanda respectivă și să o activăm tastând Ctrl+R. În fișierul de script pot fi adăugate și comentarii sau explicații cu privire la comenzile folosite. Orice comentariu trebuie să aibă la început de rând simbolul #, pentru ca ceea ce este cuprins în comentariu să nu fie considerat cod.

Un fișier de tip script are avantajul că poate fi utilizat din nou pentru aceleași date sau pentru un set nou de date statistice. De asemenea, fișierul poate fi modificat prin adăugarea sau eliminarea unor comenzi.

Fereastra R Editor poate fi utilizată curent pentru a scrie comenzile necesare, iar execuția lor se poate realiza din acest loc. Avantajul acestei abordări este acela că se pot edita mult mai ușor liniile de comandă decât în consola R.

Import fișiere de date

Cel mai frecvent, dispunem de date statistice care sunt deja structurate în fișiere de tip text sau în cele de tip calcul tabelar (fișiere Excel – *.xls sau *comma-separeted values* – *.csv). Asemenea date pot fi importate în R cu ajutorul funcției *read*. În funcție de tipul fișierului - sursă și de informațiile ce le conține, sintaxa comenzii poate să difere. Pentru a vedea toate detaliile sintaxei, se poate apela la comanda > ?read.*(), situație în care trebuie specificată extensia fișierului sursă.

În cazul în care dispunem de date în fișiere de tip Excel, acestea pot fi salvate în format *.csv și apoi importate în R cu ajutorul funcției > read.csv(). În argumentul funcției, trebuie introdusă informația despre locația fișierului sursă. Comanda se scrie astfel:

```
> nume set date = read.csv("C:/folder/filename.csv")
```

În argumentul funcției mai pot apărea anumite opțiuni, dintre care unele au valoare logică (adevărat sau fals). De exemplu, dacă în fișierul sursă există o primă linie care indică numele variabilelor, atunci opțiunea *header* are valoarea *True* (care este activă implicit). De asemenea, dacă se dorește ca întreg fișierul să fie considerat o structură de date, se poate adăuga valoarea de adevăr și pentru opțiunea *data.frame* (to.data.frame = True).

Procedeul este similar dacă se dorește importarea unei structuri de date de tip SPSS. Se utilizează aceeași comandă, doar că extensia pentru tipul de fișier este *.sav. Conform bibliotecii R, comanda arată astfel:

```
(read.spss(file, use.value.labels = TRUE, to.data.frame = FALSE,
max.value.labels = Inf, trim.factor.names = FALSE,
trim_values = TRUE, reencode = NA, use.missings = to.data.frame))
```

Este important să mai precizăm că în R, pentru a importa fișiere de date, este necesar să fie activat pachetul cu numele *foreign*.

Salvare fișiere de date

Dacă se dorește să se salveze un fișier de date care au fost introduse în R, se utilizează comenzile write.table() sau write.csv(). Sintaxa este relativ simplă:

```
> write.csv(dataframe, "filename.csv")
```

BCU "ION CREANGĂ" LIBRARY

1.5.2. Analiza statistică în R

În această secțiune, prezentăm câteva detalii cu privire la posibilitățile de analiză statistică uni- și bidimensională în R. Vizăm atât analiza statistică descriptivă, cât și cea inferențială.

1. Elemente de statistică descriptivă în R

Pentru a exemplifica câteva rezultate ale analizei statistice descriptive în R, utilizăm o bază de date demografice pentru județele României, în anul 2014. Importăm datele dintr-un fișier de tip *.csv și creăm un obiect de tip structură de date după următoarea comandă:

```
> desc=read.csv("C:/Users/My work/Desktop/Demografie.csv")
```

În baza de date importată, se pot vizualiza variabilele și tipul lor aplicând comanda:

```
> sapply(desc, class)
```

Rezultatul comenzii se prezintă mai jos:

Jud	Nat	Mort	Spor	Cas	Div	Inf
"factor"	"numeric"	"numeric"	"numeric"	"numeric"	"numeric"	"numeric"

Conform tabelului de mai sus, în fișierul de date importat se găsesc 7 variabile. Pe lângă variabila categorială ce identifică județul *(Jud)*, în baza de date se găsesc 6 variabile numerice ce prezintă rate privind fenomenele demografice: natalitate, mortalitate, sporul natural, căsătorii, divorțuri, mortalitate infantilă.

Prezentăm câteva rezultate privind analiza statistică descriptivă în R pornind de la analiza grafică și, apoi, calculul indicatorilor statistici.

a. Analiza grafică

Tipologia graficelor depinde de tipul seriei de date utilizate și de scopul construirii reprezentărilor vizuale. În R, există mai multe pachete care oferă reprezentări grafice. Un pachet care oferă posibilitatea realizării unor reprezentări grafice complexe este *ggplot2*. Versiunea implicită a programului are la dispoziție funcția *plot()*, care permite realizarea multor tipuri de grafice.

Exemplificăm pentru variabila natalitate (*Nat*) o histogramă pentru a analiza forma distribuției județelor după rata brută de natalitate. În R, comanda este simplă:

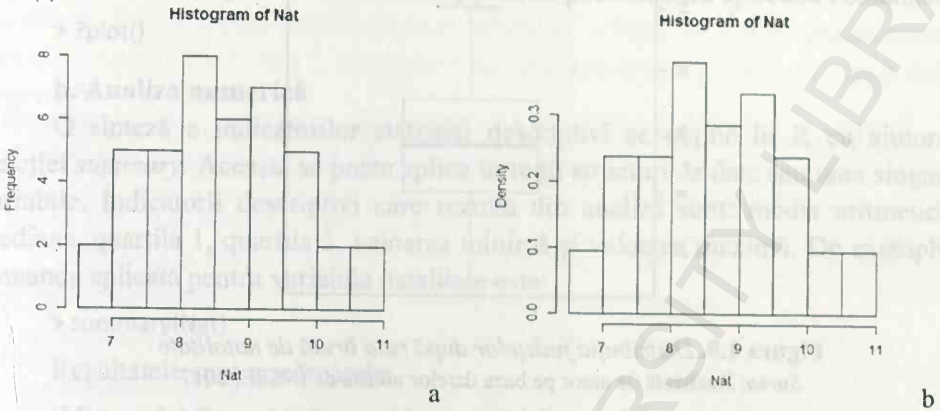
```
> hist(Nat)
```

Se obține figura 1.7 (a).

Dacă se dorește construirea unei histograme cu frecvențe relative, în sintaxa comenzii se introduce o informație legată de frecvențe:

```
> hist(Nat, freq=F)
```

Rezultă figura 1.7 (b).



a b

Figura 1.7. *Distribuţia judeţelor după rata brută de natalitate*
Sursa: Realizată de autor pe baza datelor oferite de INSSE, 2017

În fine, dacă se doreşte trasarea curbei normale corespunzătoare acestei distribuţii, se utilizează comanda:

```
> curve(dnorm(x, mean(Nat), sd(Nat)), add=T)
```

Graficul obţinut este prezentat în figura 1.8.

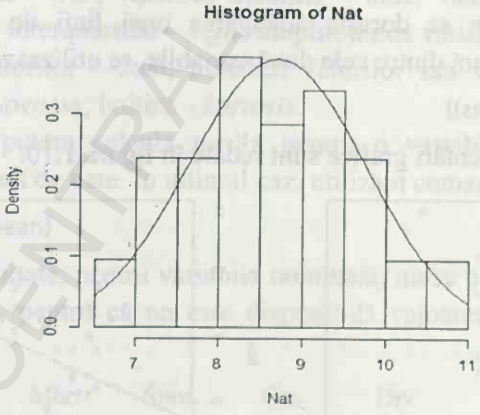


Figura 1.8. *Distribuţia judeţelor după rata brută de natalitate*
Sursa: Realizată de autor pe baza datelor oferite de INSSE, 2017

O altă reprezentare grafică importantă pentru analiza descriptivă este diagrama *box-plot*. Aceasta se construieşte cu ajutorul comenzii:

```
> boxplot(Nat)
```

BCU "ION CREANGĂ" CENTRAL UNIVERSITY LIBRARY

Rezultatul este prezentat în figura 1.9.

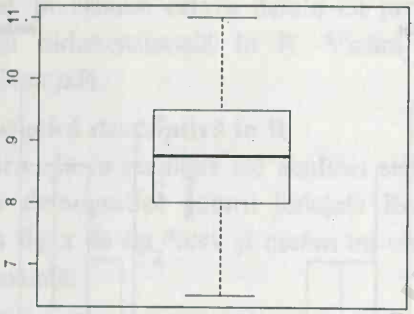


Figura 1.9. *Distribuția județelor după rata brută de natalitate*
Sursa: Realizată de autor pe baza datelor oferite de INSSE, 2017

Graficele de structură se pot construi ușor utilizând funcțiile *pie* și *barplot*.

În cazul analizei unei distribuții bidimensionale, se construiește ușor o diagramă de tip *scatter-plot*. Să considerăm exemplul relației dintre natalitate (*Nat*) și nupțialitate (*Cas*). Ordinea variabilelor în sintaxă este importantă: se precizează mai întâi variabila dependentă, urmată de simbolul „~", iar după aceasta se precizează variabila independentă. Comanda necesară este:

```
> plot(Nat~Cas)
```

În cazul în care se dorește adăugarea unei linii de regresie pentru a reprezenta legătura liniară dintre cele două variabile, se utilizează comanda:

```
> abline(lm(Nat~Cas))
```

Cele două reprezentări grafice sunt redate în figura 1.10.

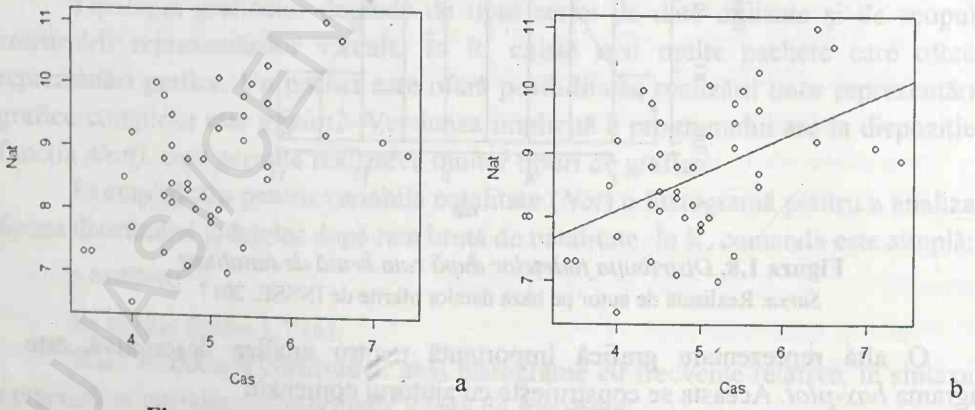


Figura 1.10. *Distribuția județelor după natalitate și nupțialitate*
Sursa: Realizată de autor pe baza datelor oferite de INSSE, 2017

BCU IAȘI CENTRALĂ UNIVERSITY LIBRARY

Pentru o reprezentare grafică, R permite o serie de opțiuni, care apar ca argumente în funcția *plot*. Toate aceste opțiuni se pot vizualiza aplicând comanda:

```
> ?plot()
```

b. Analiza numerică

O sinteză a indicatorilor statistici descriptivi se obține în R cu ajutorul funcției *summary*. Aceasta se poate aplica întregii structuri de date sau unei singure variabile. Indicatorii descriptivi care rezultă din analiză sunt: media aritmetică, mediana, quartila 1, quartila 3, valoarea minimă și valoarea maximă. De exemplu, comanda aplicată pentru variabila natalitate este:

```
> summary(Nat)
```

Rezultatele sunt următoarele:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6,500	8,000	8,750	8,695	9,450	11,000

Dacă în baza de date există date lipsă (*non-available* sau *na*), în sintaxă se adaugă un operator logic care să elimine (*remove*) aceste valori din analiza statistică. În argumentul funcției, se utilizează opțiunea: *na.rm=T*.

În R, pentru fiecare indicator statistic descriptiv, există o funcție corespunzătoare: medie – *mean*; mediană – *median*; abaterea medie pătratică – *sd*; dispersia sau varianța – *var*; valoarea maximă – *max*; valoarea minimă – *min*; mărimea intervalului interquartilic – *IQR*; amplitudinea variației – *range*; quantile – *quantile*; suma valorilor – *sum*; numărul valorilor sau volumul populației – *length*; asimetrie – *skewness*; boltire – *kurtosis*.

De exemplu, putem calcula media pentru o variabilă sau pentru toate variabilele din structura de date. În ultimul caz, utilizăm comanda:

```
> sapply(desc, mean)
```

În setul de rezultate, pentru variabila nominală, nu se obține o valoare, ci se returnează codul *NA*, pentru că nu este disponibilă valoarea cerută prin funcția medie:

Jud	Nat	Mort	Spor	Cas	Div	Inf
NA	8,695	11,735	-3,040	5,009	1,185	8,911

În cazul seriilor bidimensionale, se pot obține o serie de indicatori care analizează relația dintre două variabile. Pentru cazul variabilelor numerice, se pot determina indicatorii (cu funcțiile menționate): covarianța – *cov*; coeficientul de corelație Pearson – *cor*; coeficientul de corelație Spearman – *cov (var1, var2, method="spearman")*. Indicatorii se pot calcula și pentru întreaga structură de date, obținându-se o matrice de rezultate corespunzătoare.

BCU CENTRALA UNIVERSITARA LIBRARY

De exemplu, pentru variabilele natalitate și nupțialitate, se poate determina coeficientul de corelație Pearson. Comanda și rezultatul sunt:

```
> cor(Nat, Cas)
[1] 0,482
```

Pentru a construi matricea coeficienților de corelație, este necesar să eliminăm prima coloană din structura de date (obiectului care prezintă structura de date i se asociază argumentul [-1]), pentru că este o variabilă nominală. Comanda și rezultatul sunt prezentate mai jos:

```
> cor(desc[-1])
             Nat       Mort       Spor        Cas        Div        Inf.
Nat    1.0000000 -0.5718455  0.83197259  0.4829942 -0.18681707 -0.23753044
Mort  -0.5718455  1.0000000 -0.93090997 -0.6003088 -0.21141653  0.49303602
Spor   0.8319726 -0.9309100  1.00000000  0.6210365  0.05980625 -0.43919953
Cas    0.4829942 -0.6003088  0.62103653  1.0000000  0.14118201 -0.57688783
Div   -0.1868171 -0.2114165  0.05980625  0.1411820  1.00000000  0.02497917
Inf.  -0.2375304  0.4930360 -0.43919953 -0.5768878  0.02497917  1.00000000
>
```

2. Elemente de statistică inferențială în R

Dintre aplicațiile statisticii inferențiale, prezentăm câteva elemente legate de estimare și testare. În R, rezultatele privind estimarea și testarea se obțin în funcție de parametrii și obiectivele pentru care se realizează analiza. Prezentăm câteva exemple în acest sens.

a. Estimarea și testarea mediei

În cazul parametrului medie, estimarea prin interval de încredere se obține ușor cu ajutorul unui funcții pentru testul Student, în care valoarea cu care se compară media este zero. În R, comanda și rezultatele pentru variabila natalitate sunt:

```
> t.test(Nat)

One Sample t-test
data: Nat
t = 53,318,   df = 41,   p-value < 2,2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
8,365883;   9,024593
sample estimates:
mean of x: 8,695238
```

Așa cum se observă din rezultatele de mai sus, opțiunea implicită pentru nivelul de încredere este de 0,95. Acesta se poate modifica printr-o opțiune suplimentară în sintaxă. De exemplu, dacă se dorește o probabilitate de 0,9, la

argumentul funcției se adaugă opțiunea *conf.level=0,9*. De asemenea, în funcția de testare este considerat implicit testul bilateral, dar se poate construi și un test unilateral. De exemplu, pentru un test unilateral stânga, la argument se adaugă opțiunea *alternative="less"*, iar pentru un test unilateral dreapta, opțiunea este: *alternative="greater"*.

Aceeași funcție se utilizează pentru testarea egalității a două medii, la argument indicându-se cele două variabile pentru care se dorește testarea.

b. Testarea coeficientului de corelație

Considerăm exemplul de mai sus pentru variabilele natalitate și nupțialitate. Testarea coeficientului de corelație Pearson presupune comanda:

```
> cor.test(Nat,Cas)
```

Rezultatele sunt următoarele:

```
Pearson's product-moment correlation
data:  Nat and Cas
t = 3,4886,    df = 40,    p-value = 0,001196
alternative hypothesis:  true mean is not equal to 0
95 percent confidence interval:
0,2098713;   0,6861942
sample estimates:
cor:  0,4829942
```

c. Teste cu privire la legea de distribuție a unei variabile

Cel mai frecvent se verifică ipoteza de normalitate a unei distribuții. În acest scop, se poate utiliza abordarea grafică sau cea numerică.

Grafic, testarea normalității se poate realiza cu ajutorul diagramei Q-Q Plot. Comanda este simplă:

```
> qqnorm(Nat) – pentru a vizualiza frecvențele empirice și teoretice;
> qqline(Nat) – pentru a vizualiza linia specifică distribuției normale.
```

Rezultă diagrama din figura 1.11.

În testele numerice, ipoteza nulă presupune că variabila admite o lege de distribuție normală. Un test numeric frecvent utilizat pentru a verifica normalitatea unei distribuții este testul Shapiro. Comanda în R împreună cu rezultatul sunt prezentate mai jos:

```
> shapiro.test(Nat)

Shapiro-Wilk normality test
data:  Nat
W = 0,99028,    p-value = 0,9738
```

BCU "OVIDIUS" CONSTANȚA UNIVERSITY LIBRARY

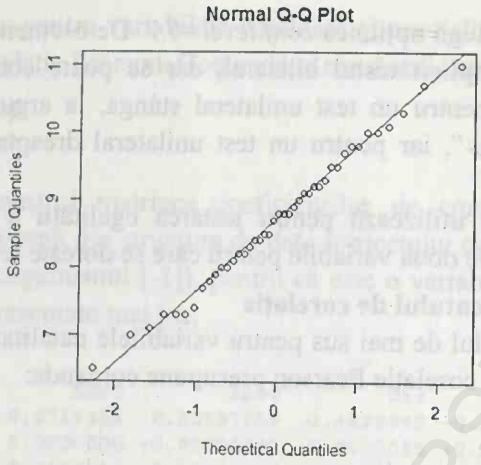


Figura 1.11. *Diagrama Q-Q Plot pentru natalitate*
Sursa: Realizată de autor pe baza datelor oferite de INSSE, 2017

Un alt test des utilizat în analiza econometrică este testul Jarque-Bera (test ce poate fi utilizat din pachetul *tseries*). Comanda în R și rezultatele sunt:

```
> jarque.bera.test(Nat)

Jarque Bera Test
data: Nat
X-squared = 0,63528, df = 2, p-value = 0,7279
```

Conform rezultatelor, în cazul ambelor teste, nu există motive să se respingă ipoteza de normalitate.

Observație

Codul R utilizat în această secțiune pentru a exemplifica analiza statistică poate fi salvat într-un fișier script și poate fi reutilizat. Putem utiliza acest cod pentru două noi variabile economice la nivelul județelor României, în anul 2014: rata șomajului și nivelul PIB pe locuitor (lei). Liniile de cod sunt prezentate mai jos și pot fi privite ca un întreg, ca o rutină ce poate fi aplicată ca un întreg. Pentru a reprezenta cele 6 grafice într-o singură fereastră, se aplică comanda par(mfrow=c(2,3)), ce așează graficele pe 2 linii și 3 coloane.

```
econ<-read.csv("C:/Users/My work/Desktop/economic.csv")
attach(econ)
sapply(econ, class)
par(mfrow=c(2,3))
hist(pib, freq=F)
hist(som, freq=F)
```

```
curve(dnorm(x, mean(som), sd(som)), add=T)
boxplot(som)
boxplot(pib)
plot(som~pib)
abline(lm(som~pib))
qqnorm(som)
qqline(som)
sapply(econ, mean)
summary(pib)
cor(som,pib)
t.test(som)
cor.test(som,pib)
shapiro.test(som)
jarque.bera.test(som)
```

1.6. Teste şi exerciţii

La finalul acestui capitol, prezentăm o serie de teste grilă şi exerciţii propuse pentru rezolvare.

Teste grilă

1. Modelul econometric este utilizat pentru:
a) a descrie fenomenele economice;
b) a explica dependenţa dintre fenomenele economice;
c) a realiza predicţii asupra realităţii economice supuse analizei.

2. Parametrii modelului econometric sunt:
a) variabile aleatoare;
b) constante reale necunoscute;
c) valori reale calculate pe baza datelor empirice.

3. Econometria asigură suportul pentru:
a) testarea teoriilor economice;
b) dezvoltarea de noi teorii;
c) luarea deciziilor economice;
d) descrierea unui fenomen real.

4. În modelarea econometrică, se utilizează următoarele tipuri de variabile:
a) aleatoare;
b) deterministe;
c) mixte.

BCU CLUJ / UNIVERSITY LIBRARY

5. Teoria consumului formulată de Keynes presupune:
a) o relație de dependență liniară între venit și consum;
b) că între venit și consum există o legătură directă;
c) venitul și consumul cresc în aceeași măsură;
d) creșterea consumului se realizează proporțional cu creșterea venitului.

6. Conceptul de model vizează:
a) o reproducere fidelă a realității;
b) un instrument de descriere a unui fenomen;
c) o reprezentare schematică a realității.

7. Cu ajutorul econometriei se poate studia:
a) legătura dintre preț și cerere;
b) evoluția în timp a prețului unei acțiuni;
c) variația fertilității în timp și la nivel de regiuni sub impactul creșterii economice.

Exerciții

1. În cazul teoriei consumului formulată de Keynes, prezentați diferența dintre variabilele economice și cele statistice în contextul construirii unui model econometric ce verifică această teorie la nivelul economiei românești.

2. În cazul teoriei lui Philips privind relația dintre șomaj și inflație, specificați: variabila dependentă și cea independentă; tipul legăturii dintre variabile (liniară sau neliniară) și justificarea unei anumite opțiuni; ce ecuație matematică ar reprezenta cel mai bine legătura dintre cele două variabile.

3. Se analizează legătura dintre cererea de benzină și prețul acestui produs pentru o perioadă de timp. Prezentați etapele demersului metodologic al construirii unui model econometric ce explică legătura dintre cele două variabile economice.

4. Utilizând R, scrieți comanda necesară construirii unei variabile nominale care conține numele celor 8 regiuni de dezvoltare ale României. Adăugați o nouă variabilă care se referă la rata totală de fertilitate regională în anul 2014, utilizând sursa de date www.insse.ro. Cu cele două variabile, creați o structură de date.

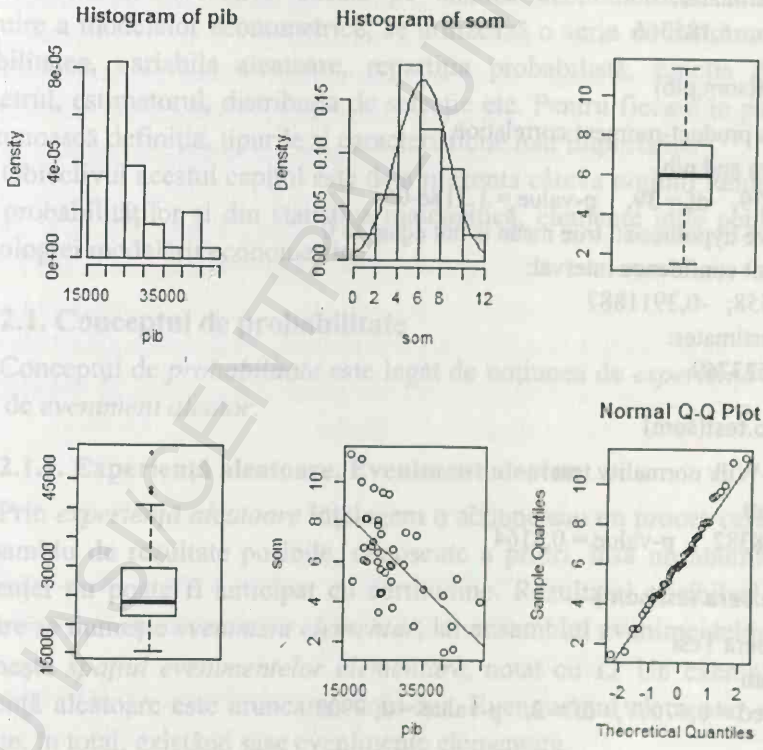
5. Pentru structura de date creată la exercițiul 4, realizați o analiză descriptivă ce presupune analiza grafică și cea numerică. Construiți un script pentru analiza realizată. Comentați rezultatele obținute.

BCU ASE BUCHAREST UNIVERSITY LIBRARY

6. Interpretați rezultatele de mai jos, obținute cu ajutorul codului prezentat la finalul secțiunii anterioare pentru rata șomajului și nivelul PIB pe locuitor (lei), la nivelul județelor României, în anul 2014.

```
> sapply(econ, class)
      pib       som
"integer"  "numeric"

> par(mfrow=c(2,3))
> hist(pib, freq=F)
> hist(som, freq=F)
> curve(dnorm(x, mean(som), sd(som)), add=T)
> boxplot(som)
> boxplot(pib)
> plot(som~pib)
> abline(lm(som~pib))
> qqnorm(som)
> qqline(som)
```



BCU INVENTAR UNIVERSITY LIBRARY

```
> sapply(econ, mean)
      pib       som
26580,390244  6,185366

> summary(pib)
   Min.   1st Qu.   Median    Mean   3rd Qu.    Max.
  15440    21480    23770   26580    29700   49560

> cor(som,pib)
[1] -0,623769

> t.test(som)

One Sample t-test
data:  som
t = 16,688,   df = 40,   p-value < 2,2e-16
alternative hypothesis:  true mean is not equal to 0
95 percent confidence interval:
5,436278;    6,934454
sample estimates:
mean of x:  6,185366

> cor.test(som,pib)

Pearson's product-moment correlation
data:  som and pib
t = -4,9839,   df = 39,   p-value = 1,318e-05
alternative hypothesis:  true mean is not equal to 0
95 percent confidence interval:
-0,7814558;  -0,3911887
sample estimates:
cor:  -0,623769

> shapiro.test(som)

Shapiro-Wilk normality test
data:  som
W = 0,98382,    p-value = 0,8164

> jarque.bera.test(som)

Jarque Bera Test
data:  som
X-squared = 0,47077,   df = 2,   p-value = 0, 7903
```

BCU IASI CENTRAL UNIVERSITY LIBRARY

Capitolul 2

ELEMENTE DE TEORIA PROBABILITĂŢILOR ŞI DE STATISTICĂ MATEMATICĂ

Elementele de teoria probabilităţilor şi de statistică matematică sunt indispensabile în cercetarea bazată pe analiza econometrică. În procesul de construire a modelelor econometrice, se utilizează o serie de instrumente precum: probabilitatea, variabila aleatoare, repartiţia probabilistă, funcţia de repartiţie, parametrul, estimatorul, distribuţia de selecţie etc. Pentru fiecare în parte, este util să se cunoască definiţia, tipurile şi caracteristicile mai importante.

Obiectivul acestui capitol este de a prezenta câteva noţiuni fundamentale din teoria probabilităţilor şi din statistica matematică, elemente utile pentru aplicarea metodologiei modelării econometrice.

2.1. Conceptul de probabilitate

Conceptul de *probabilitate* este legat de noţiunea de *experienţă aleatoare* şi de cea de *eveniment aleator*.

2.1.1. Experienţă aleatoare. Eveniment aleator

Prin *experienţă aleatoare* înţelegem o acţiune sau un proces care conduce la un ansamblu de rezultate posibile, cunoscute a priori, însă un anumit rezultat al experienţei nu poate fi anticipat cu certitudine. Rezultatul posibil al experienţei aleatoare se numeşte *eveniment elementar*, iar ansamblul evenimentelor elementare se numeşte *spaţiul evenimentelor elementare*, notat cu Ω. Un exemplu clasic de experienţă aleatoare este aruncarea unui zar. Evenimentul elementar este apariţia unei feţe, în total, existând şase evenimente elementare.

BCU CLUJ-NAPOCA UNIVERSITY LIBRARY

Unei experienţe aleatoare i se asociază un spaţiu Ω ale cărui elemente, $\omega_1, \omega_2, ..., \omega_n, ...$, reprezintă rezultatele posibile ale experienţei sau evenimentele elementare. Spaţiul Ω poate fi finit sau infinit.

Conform exemplului de mai sus, al experienței aruncării unui zar, mulţimea evenimentelor elementare este: $\Omega = \{1,2,3,4,5,6\}$, care este o mulţime finită.

Un eveniment definit printr-o proprietate, care poate fi îndeplinită sau nu în urma realizării experienţei aleatoare, se numeşte *eveniment aleator*. Evenimentul aleator este definit de un subansamblu sau de o submulţime de evenimente elementare care posedă proprietatea cerută. De exemplu, în cazul aruncării zarului, un eveniment aleator îl constituie apariţia unei feţe cu număr par. Submulţimea care corespunde acestei proprietăţi este *{2, 4, 6}*. Evenimentele care compun această submulţime se numesc *evenimente favorabile* realizării evenimentului aleator definit prin proprietatea respectivă. Un eveniment aleator este realizat dacă rezultatul experienţei aleatoare este un element din submulţimea evenimentelor elementare favorabile.

Asupra evenimentelor aleatoare se pot aplica relaţiile şi operaţiile logice specifice teoriei mulţimilor: relaţii – implicaţie şi egalitate; operaţii – reuniune, intersecţie şi complementaritate.

Două evenimente aleatoare se numesc *incompatibile* sau *mutual exclusive* într-o experienţă dacă nu pot fi realizate simultan, adică dacă nu au nici un eveniment elementar comun. Evenimentele A şi B sunt incompatibile dacă:

$A \cap B = \varnothing$.

De exemplu, în cazul aruncării unui zar, evenimentul apariţiei unui număr mai mare decât 4 şi cel al apariţiei unui număr mai mic sau egal cu 3 sunt incompatibile.

O clasă de subansamble sau părţi nevide ale lui Ω, notată $\{P_j, j \in J\}$, unde J este o mulţime cel mult numărabilă, se numeşte *partiţie* sau *sistem de evenimente mutual exclusive şi exhaustive* dacă au loc condiţiile:

- $P_i \cap P_j = \Phi, \ i \neq j, \ \ i,j \in J$;
- $\bigcup_{j \in J} P_j = \Omega$.

Dacă sistemul de evenimente definit mai sus face parte dintr-un corp de evenimente K din Ω, atunci acesta poartă numele de *sistem complet de evenimente* sau *sistem exhaustiv de evenimente*.

Un corp de evenimente K se mai numeşte şi *algebră* şi se defineşte prin următoarele proprietăţi:

- $(\forall) \ A \in K \Rightarrow \ A^c \in K$;

- $(\forall)\ A, B \in K \Rightarrow A \cup B \in K$;
- $\Omega \in K$.

Dubletul $\{\Omega, K\}$ se numeşte *câmp de evenimente*. Acesta este dat de mulţimea evenimentelor elementare Ω şi de un corp de evenimente K definit pe Ω.

2.1.2. Definiţia probabilităţii

Definirea conceptului de *probabilitate* a suferit modificări de-a lungul timpului. Există două accepţiuni ale termenului de *probabilitate*: probabilitatea subiectivă şi probabilitatea obiectivă.

Probabilitatea subiectivă

Într-o primă formă, probabilitatea subiectivă implică o conotaţie psihologică, fiind o expresie a gradului de certitudine sau incertitudine cu privire la valoarea de adevăr a unor propoziţii. Mai apoi, probabilitatea subiectivă primeşte şi o accepţiune logică, ca o expresie a gradului de încredere raţională, care are la bază anumite cunoştinţe cu privire la valoarea de adevăr a unor propoziţii.

Probabilitatea obiectivă

În această accepţiune, un enunţ de probabilitate este tratat ca un enunţ[1] despre frecvenţa relativă a anumitor evenimente în cadrul unui şir de evenimente. Pentru probabilitatea obiectivă, se cunosc mai multe definiţii, care corespund etapelor de evoluţie a conceptului în sine.

a. Definiţia clasică a probabilităţii este atribuită lui Bernoulli şi Laplace[2]. Această definiţie poate fi rezumată astfel: probabilitatea ca un eveniment să se realizeze reprezintă raportul dintre numărul de evenimente elementare favorabile realizării evenimentului şi numărul evenimentelor egal posibile.

Aşadar, probabilitatea este un raport dintre cardinalul a două mulţimi: cardinalul mulţimii evenimentelor favorabile şi cardinalul mulţimii evenimentelor egal posibile Ω. Formalizând, avem relaţia:

$$p = \frac{card(A)}{card(\Omega)} = \frac{m}{n},$$

unde m este numărul cazurilor favorabile şi n este numărul cazurilor posibile.

Între m şi n există relaţia $0 \le m \le n$, ceea ce implică $0 \le p \le 1$. Valoarea $p = 0$ corespunde imposibilităţii realizării evenimentului sau corespunde

---

[1] K. Popper, *Logica cercetării*, Editura Ştiinţifică şi Enciclopedică, Bucureşti, 1981, p. 167.

[2] P.S. Laplace, *Théorie analytique des probabilités*. Paris, Courcier, 1812, Introd., p. vii.

evenimentului *imposibil*, iar valoarea $p=1$ corespunde evenimentului *cert* sau sigur. Evenimentul sigur este dat de mulţimea evenimentelor elementare Ω. Probabilitatea obţinută prin această definiţie a fost numită *probabilitate a priori*.

Definiţia clasică a fost criticată de specialişti (de exemplu, Poincaré), pentru că presupune „echiprobabilitatea" evenimentelor care sunt implicate într-o experienţă, ceea ce reprezintă o limită sau un caz particular.

b. Definiţia probabilităţii bazate pe frecvenţă încearcă să înlăture critica adusă definiţiei clasice. Noţiunea de *frecvenţă* este o noţiune statistică care se bazează pe observarea şi înregistrarea numărului efectiv de realizări m ale unui eveniment dintr-un număr n de experienţe realizate. Probabilitatea este definită ca un caz limită al frecvenţei, atunci când numărul de experienţe tinde la infinit. Prin formalizare, se poate scrie:

$$p = \lim_{n\to\infty} \frac{m}{n}.$$

Această definiţie arată că, repetând experienţa de un număr foarte mare de ori, frecvenţa de apariţie a unui eveniment se stabilizează în jurul unei valori limită care reprezintă probabilitatea de apariţie a evenimentului. Probabilitatea obţinută pe baza frecvenţei se numeşte *probabilitate a posteriori*.

Această definiţie este o expresie a Legii numerelor mari (Bernoulli), care arată că frecvenţele de apariţie a unui eveniment converg către aceeaşi valoare limită când numărul de experienţe creşte. Critica adusă acestui mod de definire a probabilităţii vizează aspectul repetitiv şi de volum mare al experienţei aleatoare.

c. Definiţia axiomatică consideră că probabilitatea este o aplicaţie, o funcţie care asociază fiecărui eveniment dintr-un câmp de evenimente un număr real care măsoară posibilitatea ca evenimentul să apară prin realizarea experienţei aleatoare. Astfel, pe câmpul de evenimente $\{\Omega, K\}$, probabilitatea reprezintă funcţia $P: K \to R$, care îndeplineşte axiomele:

- $(\forall)\ A \in K,\ P(A) \geq 0$;
- $P(\Omega) = 1$;
- $(\forall)\ A,B \in K,\ A \cap B = \Phi$, *atunci* $P(A \cup B) = P(A) + P(B)$ sau

$(\forall)\ \{A_i, i = \overline{1,n}\} \subset K,\ A_i \cap A_j = \Phi$, *pentru* $i \neq j$, *atunci* $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$.

Tripletul $\{\Omega, K, P\}$ se numeşte *câmp de probabilitate*.

Pe baza definiţiei de mai sus, se pot demonstra uşor următoarele proprietăţi ale probabilităţii:

- $(\forall)\ A \in K,\ P(A^c) = 1 - P(A)$;

- $P(\Phi) = 0$;
- $(\forall)\ A \in K,\ 0 \le P(A) \le 1$;
- $(\forall)\ A, B \in K,\ P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Probabilitatea condiţionată

Există situaţii practice în care producerea unui eveniment este condiţionată de producerea altui eveniment. Pentru aceste situaţii, s-a definit *probabilitatea condiţionată.*

Pe un câmp de probabilitate $\{\Omega, K, P\}$, considerând evenimentele A, B şi $P(A) > 0$, se defineşte probabilitatea condiţionată prin funcţia $P_A : K \to R$, astfel încât:

$$P_A(B) = \frac{P(A \cap B)}{P(A)}.$$

Probabilitatea condiţionată este o probabilitate în sensul definiţiei axiomatice şi îndeplineşte proprietăţile prezentate mai sus. Probabilitatea condiţionată $P_A(B)$ are la bază o informaţie suplimentară care este oferită de realizarea evenimentului A.

Din relaţia probabilităţii condiţionate, se deduce relaţia probabilităţii intersecţiei a două evenimente:

$$P(A \cap B) = P(A) \cdot P_A(B) = P(B) \cdot P_B(A).$$

Evenimente independente

În momentul în care $P_A(B) = P(B)$, adică probabilitatea realizării evenimentului B nu depinde de informaţia obţinută prin realizarea evenimentului A, se afirmă că evenimentul B este *independent stocastic* de evenimentul A.

Două evenimente A şi B se numesc *independente stocastic* dacă are loc relaţia:

$$P(A \cap B) = P(A) \cdot P(B).$$

De exemplu, la aruncarea a două zaruri, evenimentul apariţiei cuplului (6;6) este un eveniment compus din două evenimente elementare independente. Rezultă că probabilitatea evenimentului cerut se obţine ca produs al probabilităţilor evenimentelor individuale. Rezultă că probabilitatea cerută este 1/36.

2.2. Variabile aleatoare

Instrumentul matematic fundamental utilizat în statistica inferenţială şi în econometrie este variabila aleatoare. Prezentăm succint câteva elemente legate de definirea şi proprietăţile variabilelor aleatoare.

2.2.1. Definiţie şi tipuri de variabile aleatoare

O experienţă aleatoare este descrisă prin mulţimea evenimentelor elementare $\Omega = \{\omega_1, \omega_2, ...\omega_n, ...\}$. Această mulţime poate fi finită sau infinită.

Definirea variabilei aleatoare are la bază ideea asocierii unui număr real fiecărui eveniment elementar ω, printr-o operaţie de măsurare. Aşadar, realizăm corespondenţa:

$$\omega_i \to X(\omega_i) = x_i, \quad x_i \in R,$$

unde X este o funcţie definită pe Ω, cu valori în mulţimea numerelor reale: $X : \Omega \to R$.

Acestei funcţii i se impune condiţia de măsurabilitate, deoarece fiecărui eveniment elementar i se asociază un număr real, printr-o operaţie de măsurare. Funcţiile reale de acest gen poartă numele de *variabile aleatoare*.

Pe câmpul de probabilitate $\{\Omega, K, P\}$, variabila aleatoare este o funcţie măsurabilă, $X : \Omega \to R$.

Un exemplu de variabilă aleatoare este cea asociată experienţei aleatoare a aruncării a două zaruri. În acest caz, $\Omega = \{(i, j), i, j = \overline{1,6}\}$, iar funcţia care se poate asocia experienţei este aceea a atribuirii unui număr real fiecărui eveniment elementar egal cu suma punctelor obţinute la fiecare aruncare. Deci,

$\omega = (i, j)$, iar $X(\omega) = i + j$, pentru $i, j = \overline{1,6}$.

Pentru a prezenta câteva dintre proprietăţile unei variabile aleatoare, ne vom folosi de o clasă specială de variabile aleatoare, numite *variabile aleatoare simple*. Pe un câmp de probabilitate $\{\Omega, K, P\}$, o asemenea variabilă ia valori distincte şi are o mulţime a valorilor finită:

$X(\Omega) = \{x_1, x_2, ...x_n\}$, cu $x_i \neq x_j \in R$.

Evenimentele $A_i = \{\omega \in \Omega,\ X(\omega) = x_i, i = \overline{1,n}\} \in K$ formează *un sistem complet de evenimente*, adică:

$$A_i \neq A_j, i \neq j \Rightarrow A_i \cap A_j = \Phi, \quad i, j = \overline{1,n} \text{ şi } \bigcup_{i=1}^{n} A_i = \Omega.$$

În aceste condiţii, notând cu $p_i = P(A_i)$, rezultă:

$$\sum_{i=1}^{n} p_i = 1.$$

Deoarece $A_i = \{\omega \in \Omega,\ X(\omega) = x_i, i = \overline{1,n}\} \in K$, se poate vorbi despre probabilitatea ca variabila aleatoare X să ia o anumită valoare, adică:

$$p_i = P(A_i) = P(X = x_i) = P\{\omega \in \Omega,\ X(\omega) = x_i,\ i = \overline{1,n}\}.$$

BCU IASI / CENTRAL UNIVERSITY LIBRARY

Probabilităţile p_i împreună cu valorile x_i determină o repartiţie de probabilitate pentru variabila aleatoare X, care se poate scrie astfel:

$$X:\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i=\overline{1,n}}.$$

Funcţia de repartiţie

Repartiţia sau *legea de probabilitate* a unei variabile aleatoare este dată prin funcţia sa de probabilitate $P^{(X)}$, care poate lua diferite forme, în funcţie de tipul variabilei. Pe baza funcţiei de probabilitate a unei variabile aleatoare, se determină funcţia sa de repartiţie.

În general, *funcţia de repartiţie* este definită prin relaţia:

$F(x)=P^{(X)}((-\infty,x)),\ (\forall)\ x\in R$ sau

$F(x)=P^{(X)}[X^{-1}((-\infty,x))]=P(\{\omega\in\Omega,\ X(\omega)<x\}),\ (\forall)\ x\in R.$

Simplificat, se mai poate scrie:

$F(x)=P(X<x),\ (\forall)\ x\in R.$

Funcţia de repartiţie are următoarele proprietăţi:

- $(\forall)\ x\in R,\ 0\le F(x)\le 1;$
- $(\forall)\ a,b\in R,\ a<b,\ F(a)\le F(b);$
- $\lim\limits_{x\to+\infty}F(x)=1;$
- $\lim\limits_{x\to-\infty}F(x)=0;$
- $(\forall)\ x\in R,\ F(x-0)=F(x).$

Pentru o variabilă aleatoare X şi pentru oricare două constante reale a, b, care îndeplinesc condiţia $a < b$, se poate demonstra următorul rezultat:

$P(a\le X<b)=F(b)-F(a),\ (\forall)\ x\in R.$

Variabile aleatoare discrete şi continue

Variabila aleatoare simplă, folosită ca exemplu până aici, face parte din familia variabilelor aleatoare discrete. În general, *o variabilă aleatoare discretă* ia valori distincte pe o mulţime a valorilor sale I, care este o mulţime cel mult numărabilă.

Matematic, variabila aleatoare discretă este definită prin:

$$X:\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}.$$

Pentru această variabilă, funcţia de repartiţie este:

BCU CENTRAL UNIVERSITY LIBRARY

$$F(x) = \sum_{\{i: x_i < x\}} p_i .$$

O altă familie importantă a variabilelor aleatoare este cea a căror mulțime a valorilor este un interval mărginit sau nemărginit. Aceste variabile aleatoare se numesc continue. *Variabilele aleatoare continue* sunt date prin funcția lor de repartiție:

$$F(x) = \int_{-\infty}^{x} f(t) \cdot dt \,, \; (\forall)\, x \in R \,,$$

unde $f(t)$ se numește *funcție densitate de repartiție.*

Funcția densitate de repartiție are următoarele proprietăți:

- $(\forall)\; x \in R,\; f(x) \geq 0 ;$
- $\int_{-\infty}^{+\infty} f(x) \cdot dx = 1 ;$
- $(\forall)\; a, b \in R,\; a < b,\; \int_{a}^{b} f(x) \cdot dx = F(b) - F(a) ;$
- $f(x) = F'(x) .$

În inferența statistică, se folosesc o serie de repartiții probabilistice mai importante, dintre care amintim:

- repartiții discrete: Bernoulli, uniformă, binomială, hipergeometrică, geometrică, Poisson etc.;
- repartiții continue: normală (standard și generalizată), Beta, Gama, Chi-pătrat, Student, Snedecor-Fisher etc.

2.2.2. Caracteristici numerice ale unei variabile aleatoare

Distribuția unei variabile aleatoare poate fi caracterizată ușor prin intermediul unor mărimi numite *parametri* sau *caracteristici numerice*. Prezentăm mai jos câteva dintre aceste caracteristici numerice specifice unei variabile aleatoare.

Media unei variabile aleatoare

Conform cu literatura de specialitate de limbă engleză și franceză, pentru medie (valoarea așteptată, speranța matematică) utilizăm și notația $E(X)$.

O variabilă aleatoare X, definită pe câmpul $\{\Omega, K, P\}$, admite medie dacă este P – integrabilă sau P – sumabilă. În aceste condiții, media se obține prin relația:

BCU MOLDOVA UNIVERSITY LIBRARY

$$E(X)=\int_R x\cdot dF(x).$$

Dacă variabila X este discretă, atunci:

$$E(X)=\sum_{i\in I} x_i\cdot p_i,$$

iar dacă variabila X este continuă, media este dată prin relaţia:

$$E(X)=\int_R x\cdot f(x)\cdot dx.$$

Proprietăţi

- pentru orice constantă reală c, $E(c)=c$;
- $E(cX)=cE(X)$;
- în cazul a două variabile aleatoare X, Y şi două constante reale a, b, are loc: $E(aX+bY)=aE(X)+bE(Y)$;
- pentru două variabile aleatoare independente: $E(X\cdot Y)=E(X)\cdot E(Y)$.

Varianţa unei variabile aleatoare

Conform cu literatura de specialitate de limbă engleză şi franceză, pentru varianţă utilizăm şi notaţia $V(X)$.

Prin definiţie, dispersia unei variabile aleatoare este dată de relaţia:

$$V(X)=E[X-E(X)]^2=E(X^2)-E^2(X),$$

unde $E(X^2)$ reprezintă media variabilei $X^2=X\cdot X$.

Existenţa dispersiei impune condiţia ca variabila să fie pătrat sumabilă, în raport cu probabilitatea.

Proprietăţi

- pentru orice constantă reală c, $V(c)=0$;
- $V(cX)=c^2V(X)$;
- în cazul a două variabile aleatoare independente X, Y şi două constante reale a, b, are loc:

$$V(aX+bY)=a^2V(X)+b^2V(Y).$$

Abaterea medie pătratică a unei variabile aleatoare

Numită şi *abatere standard*, abaterea medie pătratică se obţine ca rădăcină pătrată a varianţei: $\sqrt{V(X)}$.

Se interpretează ca o medie şi este o măsură a gradului de dispersie la nivelul repartiţiei unei variabile aleatoare.

BCU "ION CREANGĂ" UNIVERSITY LIBRARY

Momentele unei variabile aleatoare

- *momentul inițial de ordin k* este definit prin relația:

$\alpha_k(X) = E(X^k),\ k \in N$;

- *momentul absolut de ordin k* este dat prin:

$\beta_k(X) = E(|X|^k),\ k \in R,\ k > 0$;

- *momentul centrat de ordin k* se obține prin relația:

$\mu_k(X) = E\{[X - E(X)]^k\},\ k \in N$.

Covarianța a două variabile aleatoare

Pentru două variabile aleatoare X și Y, se definește caracteristica numerică numită *covarianță* sau *corelație* prin relația:

$$\text{cov}(X,Y) = E\{[X - E(X)][Y - E(Y)]\} = E(X \cdot Y) - E(X) \cdot E(Y).$$

Proprietăți

- dacă variabilele X și Y sunt independente stocastic, atunci $\text{cov}(X,Y) = 0$, deoarece, în acest caz, $E(X \cdot Y) = E(X) \cdot E(Y)$;
- $\text{cov}(X,X) = V(X)$;
- $\text{cov}(X,Y) = \text{cov}(Y,X)$;
- $\text{cov}(X, Y+Z) = \text{cov}(X,Y) + \text{cov}(X,Z)$.

Coeficientul de corelație a două variabile aleatoare

În ipoteza că două variabile aleatoare X și Y au dispersii nenule, se definește *coeficientul de corelație* al celor două variabile prin relația:

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}},$$

care îndeplinește proprietatea: $-1 \le \rho(X,Y) \le 1$.

Acest coeficient măsoară intensitatea legăturii dintre cele două variabile și se interpretează în funcție de semnul și mărimea acestuia. Dacă $\rho > 0{,}7$ legătura dintre variabile este puternică; dacă $0{,}4 \le \rho \le 0{,}7$, intensitatea este medie, iar dacă este sub 0,4 legătura este slabă.

Asimetria și boltirea

Distribuția unei variabile aleatoare admite și două mărimi ce caracterizează forma distribuției: asimetria și boltirea.

Asimetria (*skewness*) este definită de relația:

$$Sw = \frac{E(X - \mu_X)^3}{\sigma_X^3}.$$

BCU "CAROL I" UNIVERSITY LIBRARY

$Sw = 0$, pentru o repartiţie normală, $Sw > 0$, pentru o asimetrie pozitivă şi $Sw < 0$, pentru o asimetrie negativă.

Boltirea (*kurtosis*) este definită de relaţia:

$$K = \frac{E(X - \mu_X)^4}{\sigma_X^4}.$$

$K = 3$, pentru o repartiţie normală, $K < 3$, pentru o repartiţie aplatizată (platicurtică) şi $K > 3$, pentru o repartiţie cu boltire (leptocurtică).

Observaţie

Este important de subliniat că în estimarea parametrului pentru boltire, anumite programe statistice calculează un coeficient care se compară cu zero, adică se calculează o valoare estimată pentru $K - 3$. În R, se utilizează relaţia de mai sus.

2.3. Parametri. Estimatori. Estimaţii

Teoria inferenţei statistice aduce în discuţie trei termeni fundamentali: parametrul, estimatorul şi estimaţia. Între aceşti termeni, se poate prezenta relaţia:

$$parametru \Leftrightarrow estimator \Leftrightarrow estimaţie$$

Un parametru este estimat prin intermediul unui estimator, cu ajutorul unei estimaţii. Estimaţia este valoarea calculată la nivelul unui eşantion, pe baza relaţiei de definiţie a estimatorului. Parametrul este estimat cu o anumită eroare, deci între estimaţie şi parametru există, de regulă, o diferenţă.

Parametrul reprezintă o valoare fixă şi necunoscută, numită şi *valoare reală* sau *adevărată* a unei populaţii studiate după o anumită caracteristică sau variabilă.

De exemplu, dacă notăm cu X – caracteristica studiată, cu P – populaţia de referinţă, cu N – volumul populaţiei şi cu θ – parametrul, atunci se poate scrie:

$\theta = f(x_1, x_2, ..., x_N)$,

adică parametrul, în expresia sa analitică, este definit în funcţie de valorile caracteristicii la nivelul populaţiei totale.

În general, parametrii asociaţi unei variabile aleatoare sunt caracteristicile sale numerice: media, dispersia (varianţa), abaterea medie pătratică, momentul de ordin k, momentul centrat de ordin k etc.

Estimatorul este o statistică, adică o variabilă aleatoare care este determinată de totalitatea eşantioanelor posibile de volum dat (n) care se pot extrage din populaţia de referinţă. Estimatorul, în expresia sa analitică, este definit ca o funcţie a variabilelor aleatoare de selecţie X_i, $i = \overline{1,n}$, definite în contextul procesului de selecţie. Astfel, estimatorul se poate defini prin relaţia:

BCU CLUJ UNIVERSITY LIBRARY

$\hat{\theta} = T(X_1, X_2, ..., X_n)$.

În procesul de estimare, cele mai importante proprietăţi ale estimatorilor sunt: nedeplasarea, convergenţa, eficienţa.

a. Nedeplasarea

Un estimator este nedeplasat dacă media sau speranţa matematică a acestuia este egală cu parametrul. Un estimator nedeplasat verifică relaţia:

$E(\hat{\theta}) = \theta$.

Dacă relaţia nu este respectată, atunci estimatorul este deplasat.

b. Convergenţa

Un estimator este convergent dacă, pentru un eşantion cu volum suficient de mare, şirul estimatorilor converge în probabilitate către parametru. Pentru un estimator convergent, are loc relaţia:

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \lambda) = 1, \quad (\forall)\ \lambda > 0.$$

c. Eficienţa

Estimatorul $\hat{\theta}$ este eficient dacă are dispersia sau varianţa cea mai mică dintre toţi estimatorii posibili pentru parametrul θ.

Ceea ce este foarte important în cazul unui estimator este repartiţia sa probabilistică, pe baza proprietăţilor căreia se realizează inferenţa statistică.

Estimaţia este o valoare dintre valorile posibile ale unui estimator. O estimaţie se obţine la nivelul unui eşantion extras, pe baza datelor culese, şi este o funcţie a valorilor de sondaj x_i, $i = \overline{1,n}$:

$\hat{\hat{\theta}} = h(x_1, x_2, ..., x_n)$.

Estimaţia este valoarea statistică calculată la nivelul eşantionului, pe baza căreia se estimează parametrul sau se ia o decizie în urma aplicării unui test statistic.

2.4. Repartiţii probabiliste bidimensionale

Considerăm două variabile aleatoare X şi Y. Pe baza lor, se defineşte vectorul aleator: $A:(X,Y)$.

În funcţie de tipul celor două variabile aleatoare, distingem două tipuri de vectori aleatori sau două cazuri: discret şi continuu.

Pentru cazul discret, au loc relaţiile:

BCU IASI / CENTRAL UNIVERSITY LIBRARY

$$A:\begin{pmatrix}(x_i,y_j)\\ p_{ij}\end{pmatrix}_{i\in I,\, j\in J},$$

$p_{ij}=P[(X=x_i)\cap(Y=y_j)]$, $\sum_i\sum_j p_{ij}=1$.

Pentru cazul continuu, au loc relaţiile:

$A: f(x,y)$, $f(x,y)\geq 0$,

$(\forall)\,(x,y)\in R^2$, $\int\limits_R\int\limits_R f(x,y)dxdy=1$;

$$F(x,y)=P(X<x,Y<y),\quad F(x,y)=\int\limits_{-\infty}^{x}\int\limits_{-\infty}^{y} f(s,t)dsdt;$$

$$f(x,y)=\frac{\partial^2 F}{\partial x\partial y}.$$

Repartiţii marginale

Pentru o repartiţie bidimensională, se definesc două repartiţii marginale. Acestea sunt distincte în funcţie de tipul variabilei: discret sau continuu.

Pentru două variabile discrete, au loc:

$$X:\begin{pmatrix}x_i\\ p_{i.}\end{pmatrix}_{i\in I},\ p_{i.}=\sum_j p_{ij},\quad \sum_i p_{i.}=1;$$

$$Y:\begin{pmatrix}y_j\\ p_{.j}\end{pmatrix}_{j\in J},\ p_{.j}=\sum_i p_{ij},\quad \sum_j p_{.j}=1;$$

$$E(X)=\sum_i x_i p_{i.}=\sum_i\sum_j x_i p_{ij},$$

$$E(Y)=\sum_j y_j p_{.j}=\sum_i\sum_j y_j p_{ij}.$$

Pentru două variabile continue, au loc:

$$X: f_1(x)=\int_R f(x,y)dy,\quad F_1(x)=\int\limits_{-\infty}^{x}\int\limits_{-\infty}^{\infty} f(s,t)dsdt=\int\limits_{-\infty}^{x} f_1(s)ds;$$

$$Y: f_2(y)=\int_R f(x,y)dx,\quad F_2(y)=\int\limits_{-\infty}^{y}\int\limits_{-\infty}^{\infty} f(s,t)dsdt=\int\limits_{-\infty}^{y} f_2(s)ds;$$

$$E(X)=\int_R x f_1(x)dx=\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} x f(x,y)dxdy.$$

BCU IAŞI FEDERAL UNIVERSITY LIBRARY

Repartiții condiționate

Pentru o repartiție bidimensională, se definesc următoarele repartiții condiționate.

Pentru cazul discret:

$$X/Y=y_j:\begin{pmatrix} x_i \\ \dfrac{p_{ij}}{p_{.j}} \end{pmatrix}_{i\in I},\quad P(X=x_i/Y=y_j)=\frac{p_{ij}}{p_{.j}};$$

$$Y/X=x_i:\begin{pmatrix} y_j \\ \dfrac{p_{ij}}{p_{i.}} \end{pmatrix}_{j\in J},\quad P(Y=y_j/X=x_i)=\frac{p_{ij}}{p_{i.}}.$$

Pentru cazul continuu:

$$X/Y: f(x/y)=\frac{f(x,y)}{f_2(y)},\quad f_2(y)\neq 0;$$

$$Y/X: f(y/x)=\frac{f(x,y)}{f_1(x)},\quad f_1(x)\neq 0.$$

Matricea de covarianță

Pentru două variabile aleatoare X și Y, matricea de covarianță se definește prin relația de mai jos:

$$\begin{pmatrix} V(X) & \operatorname{cov}(X,Y) \\ \operatorname{cov}(X,Y) & V(Y) \end{pmatrix}.$$

Medii condiționate

La nivelul fiecărei repartiții condiționate, se definește o medie condiționată.

Pentru cazul discret:

$$E(X/Y=y_j)=\sum_i x_i\frac{p_{ij}}{p_{.j}}=\frac{1}{p_{.j}}\sum_i x_i p_{ij}=\alpha_j,\ E(X/Y)=g(y);$$

$$E(Y/X=x_i)=\sum_j y_j\frac{p_{ij}}{p_{i.}}=\frac{1}{p_{i.}}\sum_j y_j p_{ij}=\beta_i,\ E(Y/X)=h(x).$$

Pentru cazul continuu:

$$E(X/Y)=\int_R xf(x/y)dx=\int_R x\frac{f(x,y)}{f_2(y)}dx=u(y);$$

$$E(Y/X)=\int_R yf(y/x)dy=\int_R y\frac{f(x,y)}{f_1(x)}dy=v(x).$$

BCU ASCENTRAL UNIVERSITY LIBRARY

2.5. Repartiţii probabiliste frecvent utilizate în econometrie

În această secţiune, prezentăm câteva dintre repartiţiile probabiliste, discrete şi continue, utilizate în analiza statistică şi econometrică.

A. Repartiţii discrete

Dintre repartiţiile discrete, prezentăm: repartiţia Bernoulli, binomială şi Poisson.

1. Repartiţia Bernoulli

Repartiţia Bernoulli se referă la variabilele aleatoare cele mai simple, care pot lua doar două valori posibile, în funcţie de realizarea sau nu a unei proprietăţi.

Repartiţia este dată printr-un tablou de forma:

$X: \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$, unde:

$\pi = P(X=1)$;

$1-\pi = P(X=0)$.

Dacă o variabilă aleatoare urmează o lege de repartiţie Bernoulli, aceasta se simbolizează prin: $X \sim B(\pi)$.

Parametrul π reprezintă media sau speranţa matematică a variabilei de tip Bernoulli şi reprezintă probabilitatea de a se realiza proprietatea cerută sau de a apărea evenimentul dorit, favorabil.

Parametrii acestei repartiţii sunt:

$E(X) = \pi$;

$V(X) = \pi(1-\pi)$.

Exemplu

Pentru valoarea $\pi = 0{,}2$, repartiţia Bernoulli se prezintă în figura de mai jos.

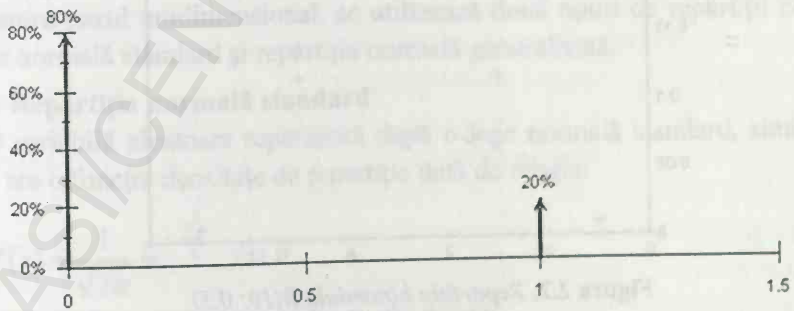


Figura 2.1. *Repartiţia Bernoulli B(0,2)*

BCU IAŞI CENTRAL UNIVERSITY LIBRARY

2. Repartiția binomială

Repartiția binomială se obține prin generalizarea repartiției Bernoulli. Prin însumarea unui număr de n variabile aleatoare Bernoulli identic repartizate, se obține o variabilă binomială.

Repartiția binomială, simbolizată $B(n,\pi)$, este dată printr-un tablou de forma:

$$X:\begin{pmatrix} k \\ C_n^k \pi^k (1-\pi)^{n-k} \end{pmatrix}_{k=\overline{0,n}},$$

unde k reprezintă numărul de realizări ale evenimentului favorabil, în condițiile repetării de n ori a experienței Bernoulli.

Parametrii repartiției binomiale se determină ușor, dacă se consideră că o variabilă binomială este suma a n variabile Bernoulli, $X_i \sim B(\pi)$.

Variabila binomială se obține astfel:

$$X = \sum_{i=1}^{n} X_i .$$

Pentru această variabilă, au loc relațiile:

$E(X) = n\pi$;

$V(X) = n\pi(1-\pi)$.

Exemplu

Pentru o experiență Bernoulli repetată de 10 ori și o probabilitate de 0,5, se obține repartiția binomială din figura 2.2.

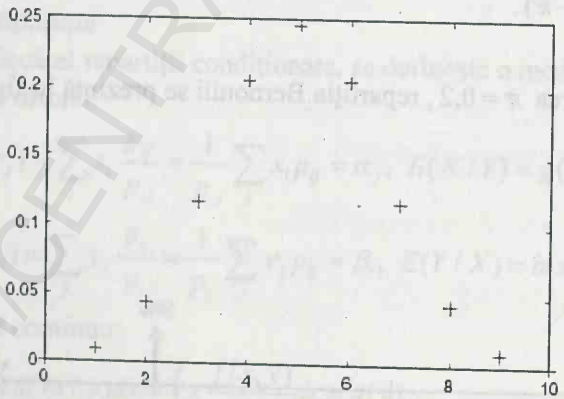


Figura 2.2. *Repartiția binomială B(10; 0,5)*

3. Repartiția Poisson

O variabilă aleatoare de tip Poisson are o lege de repartiție de forma:

BCU IASI/CENTRAL UNIVERSITY LIBRARY

$$X:\begin{pmatrix} k \\ \dfrac{e^{-\lambda}\lambda^k}{k!} \end{pmatrix}_{k=\overline{0,n}},$$

unde k reprezintă numărul de apariţii ale unui anumit eveniment într-o anumită perioadă de timp.

Specific repartiţiei Poisson, simbolizată $P(\lambda)$, este că media şi varianţa sunt egale şi au valoarea:

$E(X)=V(X)=\lambda$.

Exemplu

Pentru $\lambda=5$, se obţine o repartiţie Poisson prezentată în figura de mai jos.

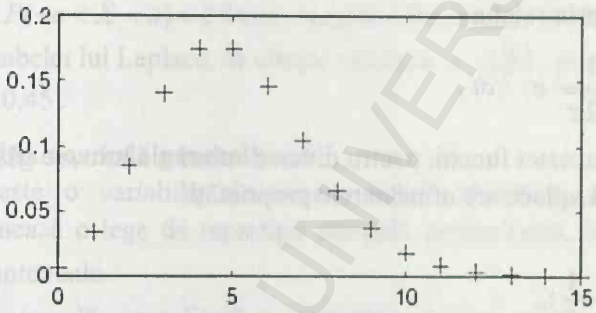


Figura 2.3. *Repartiţia Poisson P(5)*

B. Repartiţii continue

Dintre repartiţiile continue, prezentăm: repartiţia normală standard şi generalizată, Student, Chi-pătrat şi Fisher.

1. Repartiţia normală

Pentru cazul unidimensional, se utilizează două tipuri de repartiţii normale: repartiţia normală standard şi repartiţia normală generalizată.

a. Repartiţia normală standard

O variabilă aleatoare repartizată după o lege normală standard, simbolizată $N(0,1)$, are o funcţie densitate de repartiţie dată de relaţia:

$$f(x)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{x^2}{2}},\ x\in R.$$

Parametrii repartiţiei normale standard sunt:

$E(X)=0$;

$V(X)=1$.

BCU "ŞTEFAN CEL MARE" SUCEAVA UNIVERSITY LIBRARY

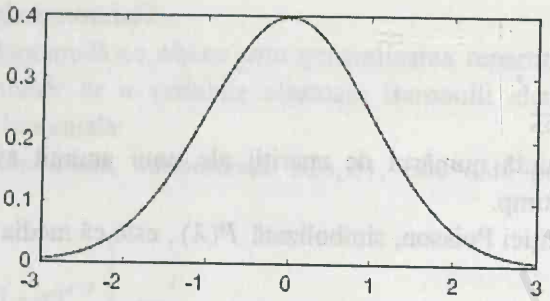


Figura 2.4. *Repartiția normală standard N(0,1)*

Pentru interese practice, de calcul al unor probabilități, se utilizează funcția lui Laplace, care este definită pe baza repartiției normale standard. Funcția lui Laplace este dată prin relația:

$$\Phi(x)=\int_0^x \frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{t^2}{2}}dt,$$

iar valorile acestei funcții, pentru diverse valori ale lui x, se găsesc tabelate.

Funcția lui Laplace are următoarele proprietăți:

- $\Phi(0)=0$;
- $\lim_{x\to+\infty}\Phi(x)=\frac{1}{2}$;
- $\Phi(-x)=-\Phi(x)$, adică este o funcție impară.

Funcția de repartiție se poate scrie și pe baza funcției lui Laplace astfel:

$$F(x)=\int_{-\infty}^{x} f(t)\cdot dt=\int_{-\infty}^{0} f(t)\cdot dt+\int_0^x f(t)\cdot dt=\frac{1}{2}+\Phi(x).$$

Pe baza valorilor funcției lui Laplace, se poate determina, de exemplu, probabilitatea ca variabila aleatoare normală standard să ia valori într-un interval simetric de tipul $(-a;a)$. Această probabilitate este:

$$P(-a<X<a)=\int_{-a}^{a} f(t)dt=F(a)-F(-a)=\frac{1}{2}+\Phi(a)-\frac{1}{2}-\Phi(-a)=2\Phi(a).$$

Exemplul 1

Pentru $a=1$, se poate calcula probabilitatea $P(-1<X<1)=2\cdot\Phi(1)=0{,}682$, deoarece din tabela lui Laplace se citește valoarea $\Phi(1)=0{,}341$.

Similar, se pot determina probabilități ca variabila normală standard să ia valori peste o anumită limită cunoscută sau sub o asemenea limită.

BCU "ION CREANGĂ" UNIVERSITY LIBRARY

Exemplul 2

Pentru $a=1$, se pot calcula probabilităţile:

$P(X<1)=F(1)=1/2+\Phi(1)=0,5+0,341=0,841$.

$P(X\geq 1)=1-P(X<1)=1-1/2-\Phi(1)=0,5-0,341=0,159$.

În practică, este important uneori să se realizeze operaţia inversă, adică, pentru o probabilitate dată, să se calculeze limita sau limitele unui interval pentru valorile variabilei aleatoare normale standard.

Exemplul 3

Pentru o probabilitate de 90%, se cere să se calculeze a, astfel încât să aibă loc:

$P(-a<X<a)=0,9$.

Deoarece $P(-a<X<a)=2\Phi(a)$, rezultă $2\Phi(a)=0,9$ şi $\Phi(a)=0,45$.

Conform tabelei lui Laplace, se citeşte valoarea $a=1,65$, pentru care:

$\Phi(1,65)=0,45$.

b. Repartiţia normală generalizată

Dacă X este o variabilă aleatoare normală standard, atunci variabila $Y=\mu+\sigma X$ urmează o lege de repartiţie normală generalizată, în care μ şi σ reprezintă constante reale.

Repartiţia normală generalizată se simbolizează $N(\mu,\sigma^2)$, unde μ şi σ^2 sunt parametrii acestei repartiţii:

$\mu=E(Y)$, respectiv $\sigma^2=V(Y)$.

În cazul unei repartiţii normale generalizate, funcţia densitate de repartiţie este dată de relaţia:

$$g(x)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Exemplul 1

O variabilă aleatoare normală generalizată de parametri *10* şi *2* este prezentată în figura 2.5.

Relaţia dintre repartiţia normală standard şi cea generalizată se poate observa uşor atât prin relaţia dintre cele două tipuri de variabile, cât şi prin relaţia existentă între funcţiile de repartiţie specifice. Astfel, unei variabile normale standard X, cu o funcţie densitate $f(x)$ şi o funcţie de repartiţie $F(x)$, îi corespunde o variabilă normală generalizată $Y=\mu+\sigma X$, cu o funcţie densitate $g(x)$ şi o funcţie de repartiţie $G(x)$, unde:

BCU CENTRAL UNIVERSITY LIBRARY

$$G(x) = F(\frac{x-\mu}{\sigma}).$$

Invers, unei variabile normale generalizate Y îi corespunde o variabilă normală standard:

$$X = \frac{Y-\mu}{\sigma}.$$

Ultima relație poartă numele de *relație de standardizare a unei variabile aleatoare.*

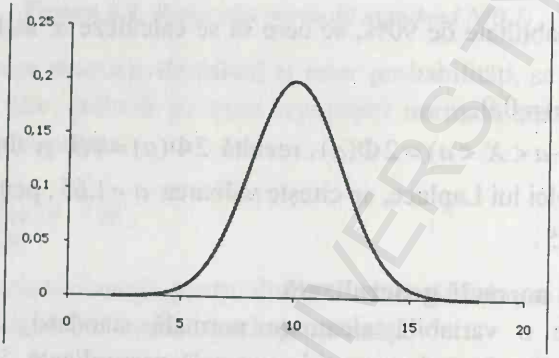


Figura 2.5. *Repartiția normală generalizată N(10,2)*

În baza operației de standardizare și a relației dintre cele două funcții de repartiție, se poate utiliza funcția lui Laplace pentru a calcula anumite probabilități pentru o repartiție normală generalizată. Astfel, se poate calcula, de exemplu, probabilitatea:

$$P(a < Y < b) = G(b) - G(a) = F(\frac{b-\mu}{\sigma}) - F(\frac{a-\mu}{\sigma}) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma}).$$

Pe baza relației de transformare de mai sus, cunoscând parametrii legii de repartiție a variabilei normale generalizate, se pot calcula diferite probabilități ca variabila să ia valori într-un anumit interval, utilizând tabela lui Laplace.

Exemplul 2

Pentru o variabilă $X \sim N(10,4)$, se cere să se calculeze probabilitatea:

$P(8 < X < 12)$ și $P(X \geq 12)$.

$$P(8 < X < 12) = \Phi(\frac{12-10}{2}) - \Phi(\frac{8-10}{2}) = \Phi(1) - \Phi(-1) =$$

$$= \Phi(1) + \Phi(1) = 2\Phi(1) = 0{,}682.$$

$$P(X \geq 12) = 1 - P(X < 12) = 1 - G(12) = 1 - F(\frac{12-10}{2}) =$$

BCU IASI CENTRAL UNIVERSITY LIBRARY

$$=1-1/2-\Phi(\frac{12-10}{2})=0,5-\Phi(1)=0,5-0,341=0,159.$$

2. Repartiţia Chi-pătrat

O variabilă aleatoare repartizată după o lege Chi-pătrat, simbolizată $\chi^2(n,\sigma)$, are o funcţie densitate de repartiţie dată prin relaţia:

$$f(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\sigma^n\Gamma\left(\dfrac{n}{2}\right)}\cdot x^{\frac{n}{2}-1}e^{-\frac{x}{2\sigma^2}}, & x>0\\ 0, & x\le 0\end{cases}$$

În relaţia de mai sus, $\Gamma(a)=\int_0^{\infty}x^{a-1}e^{-x}dx$, $a>0$, este integrala Euler.

Dacă se consideră n variabile aleatoare identic repartizate după o lege normală standard, $X_i\sim N(0,1)$, $i=\overline{1,n}$, atunci variabila:

$$X=\sum_{i=1}^{n}X_i^2\sim\chi^2(n).$$

Cu alte cuvinte, suma pătratelor a n variabile aleatoare normal repartizate reprezintă o variabilă aleatoare repartizată Chi-pătrat.

Parametrii repartiţiei Chi-pătrat depind de modul de construire a variabilei. Astfel, n reprezintă *numărul de grade de libertate*, adică numărul de variabile aleatoare normale la pătrat însumate, iar σ reprezintă abaterea standard a variabilei aleatoare normale (care este egală cu unu în cazul variabilei standard).

Caracteristicile numerice ale repartiţiei Chi-pătrat sunt:

$E(X)=n\sigma^2$;

$V(X)=2n\sigma^4$.

Exemplul 1

Pentru o variabilă $X\sim\chi^2(4)$, se poate determina constanta a, astfel încât:

$P(X>a)=0,05$.

Din tabela Chi-pătrat se citeşte valoarea $a=9,48$. Evident, prin relaţia de complementaritate specifică probabilităţilor, $P(X\le 9,48)=0,95$.

Exemplul 2

O variabilă aleatoare Chi-pătrat cu $n=4$ grade de libertate se prezintă ca în figura 2.6.

BCU "CONSTANTIN BRÂNCUŞI" UNIVERSITY LIBRARY

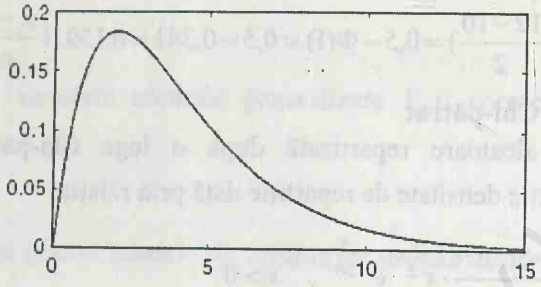


Figura 2.6. *Repartiția Chi-pătrat* $\chi^2(4)$

Pentru repartiția Chi-pătrat, sunt construite tabele probabiliste care oferă, în condițiile unui anumit grad de libertate fixat, constantele a pentru care se cunoaște: $P(X > a)$.

3. Repartiția Student

O variabilă aleatoare repartizată după o lege Student, simbolizată $t(n)$, are o funcție densitate de repartiție dată de relația:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\cdot\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in R, \text{ iar}$$

$\Gamma(a) = \int_0^{\infty} x^{a-1}e^{-x}dx, \quad a > 0$, este integrala Euler.

Dacă se consideră două variabile aleatoare $X \sim N(0,1)$ și $Y \sim \chi^2(n)$, atunci variabila aleatoare Student se obține prin relația:

$$t = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n),$$

unde n reprezintă numărul de grade de libertate, parametrul repartiției Student.

Exemplul 1

O variabilă aleatoare Student cu $n = 3$ grade de libertate se prezintă în figura 2.7.

BCU IAȘI CENTRAL UNIVERSITY LIBRARY

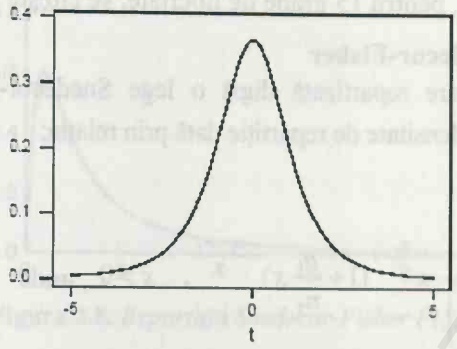


Figura 2.7. *Repartiţia Student t(3)*

În cazul repartiţiei Student, în practică, este important să se identifice anumite limite pentru care variabila ia valori într-un interval cu o probabilitate cunoscută.

Exemplul 2

Pentru o variabilă aleatoare care urmează o lege de repartiţie Student cu $n = 25$ de grade de libertate ($X \sim t(25)$), se cere să se determine valoarea a, astfel încât: $P(X > a) = 0{,}05$.

Din tabela Student, pentru o probabilitate de 0,05 şi 25 de grade de libertate, se citeşte valoarea $a = 1{,}708$.

Exemplul 3

Pentru o variabilă aleatoare $X \sim t(20)$, se cere să se determine valoarea a, astfel încât:

$P(-a < X < a) = 0{,}95$.

Simetria repartiţiei Student ne permite aflarea valorii cerute pe baza proprietăţilor:

$1 - P(-a < X < a) = 0{,}05$ şi $P(X > a) = 0{,}025$.

Din tabela Student, pentru 20 de grade de libertate, se citeşte $a = 2{,}086$.

Exemplul 4

Pentru o variabilă aleatoare $X \sim t(15)$, se cere să se determine valoarea a, astfel încât:

$P(X < -a) = 0{,}025$.

Simetria repartiţiei Student ne permite aflarea valorii cerute pe baza proprietăţii:

$P(X < -a) = P(X > a)$.

BCU "ION CREANGĂ" CENTRAL UNIVERSITY LIBRARY

Din tabela Student, pentru 15 grade de libertate, se citește $a = 2,131$.

4. Repartiția Snedecor-Fisher

O variabilă aleatoare repartizată după o lege Snedecor-Fisher, simbolizată $\Im(n_1, n_2)$, are o funcție densitate de repartiție dată prin relația:

$$f(x) = \begin{cases} \dfrac{\left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{B\left(\dfrac{n_1}{2}, \dfrac{n_2}{2}\right)} \cdot x^{\frac{n_1}{2}-1} \left(1 + \dfrac{n_1}{n_2} x\right)^{-\frac{n_1+n_2}{2}}, & x > 0, \text{ unde} \\ 0, & x \le 0 \end{cases}$$

$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$, $a > 0, b > 0$, este integrala Euler.

Caracteristicile numerice ale repartiției Snedecor-Fisher sunt:

$$E(X) = \frac{n_2}{n_2 - 2}, \quad n_2 > 2;$$

$$V(X) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}, \quad n_2 > 4.$$

Dacă se consideră două variabile aleatoare: $X \sim \chi^2(n_1, \sigma)$ și $Y \sim \chi^2(n_2, \sigma)$, atunci o variabilă repartizată Fisher se obține prin relația:

$$F = \frac{\dfrac{X}{n_1}}{\dfrac{Y}{n_2}} \sim \Im(n_1, n_2),$$

unde n_1 și n_2 reprezintă grade de libertate, parametrii repartiției Snedecor-Fisher.

Exemplul 1

O variabilă aleatoare Snedecor-Fisher, cu $n_1 = 5$ și $n_2 = 3$ grade de libertate, se prezintă ca în figura 2.8.

Pentru repartiția Snedecor-Fisher, sunt construite tabele probabiliste care oferă, în condițiile unui anumit grade de libertate, constantele a pentru care se cunoaște:

$P(X > a)$.

BCU "CAROL I" - UNIVERSITY LIBRARY

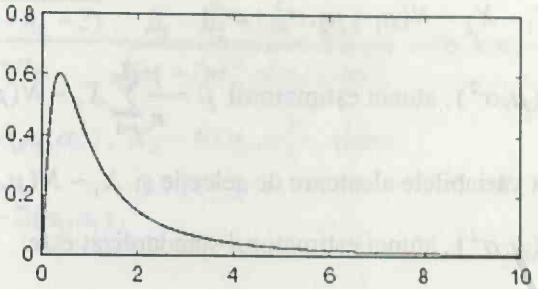


Figura 2.8. *Repartiţia Snedecor-Fisher F(5, 3)*

Exemplul 2

Pentru o variabilă $X \sim F(5,3)$, se poate determina constanta a, astfel încât: $P(X > a) = 0{,}01$.

Din tabela Fisher, se citeşte valoarea $a = 28{,}23$.

2.6. Statistici pentru eşantioane extrase din colectivităţi cu repartiţie normală

Considerăm următoarele rezultate utilizate mai des în problemele de inferenţă statistică, pentru eşantioane extrase din colectivităţi normal distribuite.

1. Dacă $Y \sim N(\mu, \sigma^2)$, atunci $X = \dfrac{Y - \mu}{\sigma} \sim N(0,1)$.

2. Dacă $X_i \sim N(0, \sigma^2)$, atunci $X = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n, \sigma)$.

3. Dacă $X_i \sim \chi^2(n_i, \sigma)$, atunci $X = \sum_{i=1}^{k} X_i \sim \chi^2(\sum_{i=1}^{k} n_i, \sigma)$.

4. Dacă $X \sim N(0, \sigma^2)$ şi $Y \sim \chi^2(n, \sigma)$, atunci

$$t = \frac{X}{\sqrt{\dfrac{Y}{n}}} \sim t(n).$$

5. Dacă $X_i \sim N(\mu_i, \sigma_i^2)$, atunci $X = \sum_{i=1}^{n} a_i X_i \sim N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)$.

BCU IAŞI CENTRAL UNIVERSITY LIBRARY

Rezultă că $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$.

6. Dacă $X \sim N(\mu, \sigma^2)$, atunci estimatorul $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$,

unde X_i sunt variabilele aleatoare de selecție și $X_i \sim N(\mu, \sigma^2)$.

7. Dacă $X \sim N(\mu, \sigma^2)$, atunci estimatorul standardizat este:

$$Z = \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

8. Dacă $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, atunci

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}).$$

9. Dacă $X_i \sim N(\mu, \sigma^2)$, atunci $Y = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n,1)$.

10. Dacă $X \sim N(\mu, \sigma^2)$, atunci $Y = \frac{(n-1)}{\sigma^2}\hat{\sigma}'^2 \sim \chi^2(n-1,1)$,

unde $\hat{\sigma}'^2$ reprezintă dispersia de selecție modificată, estimatorul nedeplasat pentru parametrul σ^2.

11. Dacă $X \sim N(\mu, \sigma^2)$, atunci $t = \frac{\hat{\mu} - \mu}{\sqrt{\frac{\hat{\sigma}'^2}{n}}} \sim t(n-1)$.

12. Dacă $X \sim N(\mu, \sigma^2)$, atunci $t = \frac{\hat{\mu} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n-1}}} \sim t(n-1)$,

unde $\hat{\sigma}^2$ este dispersia de selecție, estimatorul deplasat pentru parametrul σ^2.

13. Dacă $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, atunci

BCU IASI CENTRAL UNIVERSITY LIBRARY

$$t = \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sqrt{(n_1 - 1)\hat{\sigma}'^2_1 + (n_2 - 1)\hat{\sigma}'^2_2}} \sim t(n_1 + n_2 - 2).$$

14. Dacă $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, atunci

$$F = \frac{\sigma_2^2}{\sigma_1^2} \frac{\hat{\sigma}'^2_1}{\hat{\sigma}'^2_2} \sim \mathfrak{I}(n_1, n_2).$$

2.7. Distribuţii probabiliste în R

Utilizarea programului R permite realizarea unor calcule ce au la bază legile probabiliste cunoscute, atât discrete, cât şi continue. Astfel, cu ajutorul funcţiei de repartiţie, se poate calcula probabilitatea ca o anumită variabilă aleatoare să ia valori într-un anumit interval. De asemenea, se poate realiza demersul opus, pe baza inversei funcţiei de repartiţie, determinând limitele unui interval în care variabila ia valori cu o probabilitate cunoscută.

Pentru orice tip de distribuţie, R utilizează 4 funcţii care se bazează pe funcţia de probabilitate sau funcţia densitate şi pe funcţia de repartiţie. Cele 4 funcţii au câte un sufix specific:

- *d* se referă la funcţia densitate sau de probabilitate, pentru a calcula probabilitatea ca o variabilă aleatoare să ia o anumită valoare;
- *p* vizează funcţia de repartiţie, pentru a calcula probabilitatea ca variabila aleatoare să ia valori într-un anumit interval;
- *q* are în vedere funcţia quantilă şi foloseşte inversa funcţiei de repartiţie pentru a calcula valoarea pentru care, cu o anumită probabilitate cunoscută, variabila ia valori mai mici sau mai mari decât respectiva valoare;
- *r* se referă la funcţia care generează o variabilă aleatoare pentru o anumită lege de distribuţie.

O anumită funcţie în R se obţine adăugând la unul din prefixele de mai sus un sufix care provine de la numele repartiţiei. Câteva exemple: pentru repartiţia binomială, se foloseşte *binom*; Poisson - *pois*; Student - *t*; Fisher - *f*; Normală - *norm*; Chi-square - *chisq*; Uniformă - *unif* etc.

Deoarece legile de distribuţie sunt specificate prin intermediul unor parametri, pentru calculul probabilităţilor, este important să se precizeze valorile acestora. De exemplu, repartiţia binomială are doi parametri: parametrul π şi numărul de realizări ale evenimentului elementar *n*.

Prezentăm câteva exemple de calcul probabilistic în R, exemplificând pe câte una din cele două tipuri de distribuţii.

BCU "ASTRA" WEB LIBRARY

Exemplul 1

Se consideră experiența aruncării unei monede de un număr de 100 de ori. Să se calculeze: probabilitatea să apară stema de 30 de ori; probabilitatea ca stema să apară de mai puțin de 60 de ori; limita maximă de apariții ale stemei, cu o probabilitate de 0,9.

În acest caz, avem de a face cu legea binomială. Cei doi parametri au valorile 1/2 și 100.

Pentru cele 4 funcții prezentate mai sus, în R, sintaxa este următoarea:

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```

unde:

x, q	vector of quantiles.
p	vector of probabilities.
n	number of observations. If length(n) > 1, the length is taken to be the number required.
size	number of trials (zero or more).
prob	probability of success on each trial.
log, log.p	logical; if TRUE, probabilities p are given as log(p).
lower.tail	logical; if TRUE (default), probabilities are $P[X \leq x]$, otherwise, $P[X > x]$.

Pentru a calcula probabilitatea să apară stema de 30 de ori, se folosește comanda:

```
> dbinom(30,100,1/2)
```

Rezultatul calculului este:

```
[1] 2,317069e-05
```

Calculul celei de-a doua probabilități presupune utilizarea funcției de repartiție. Comanda în R și rezultatul sunt:

```
> pbinom(60,100,1/2)
[1] 0,9823999
```

Complementara acestei probabilități se poate calcula în două moduri:

```
> 1-pbinom(60,100,1/2)
[1] 0,0176001
```

sau

```
> pbinom(60,100,1/2, lower.tail=F)
[1] 0,0176001
```

A treia cerință presupune utilizarea comenzii:

```
> qbinom(0.9,100,1/2)
```

Rezultatul este:

```
[1] 56
```

Exemplul 2

Să considerăm cazul unei variabile cu o distribuție normală. Aceasta are doi parametri: media și abaterea standard. Sintaxa în R pentru cele 4 funcții este următoarea:

```
dnorm(x, mean = a, sd = b, log = FALSE)
pnorm(q, mean = a, sd = b, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = a, sd = b, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = a, sd = b)
```

Dacă în sintaxă nu se prezintă valorile pentru medie și abatere standard, atunci se consideră implicit cazul distribuției normale standard, cu media zero și abaterea standard egală cu 1.

Să considerăm o variabilă ce urmează legea normală standard pentru care dorim: să se calculeze probabilitatea ca variabila să ia valoarea egală cu 2 și probabilitatea ca variabila să ia valori în intervalul (-1, 1); să se determine valoarea pentru care variabila ia valori mai mici, cu o probabilitate de 0,95; să se genereze 100 de valori care urmează legea de distribuție normală standard.

Comenzile și soluțiile sunt următoarele:

```
> dnorm(2)
[1] 0,05399097
> pnorm(1)-pnorm(-1)
[1] 0,6826895
> qnorm(0.95)
[1] 1,644854
```

BCU IASI CENTRAL UNIVERSITY LIBRARY

```
> xnorm(100)
  [1]  0.103554172  0.760210805 -1.999058208 -0.387292166  0.320751635
  [6]  0.093963373  0.075234243  0.122351252  1.363760825  0.779942682
 [11]  0.734940400  0.283577306  0.839451912 -0.310185697  2.483639697
 [16] -2.223653223  1.107325047 -0.690033758  0.748953557 -0.398266960
 [21] -0.137615391  1.471872433 -0.417527534 -0.041534480  1.172192180
 [26]  0.071634493 -1.112294645 -0.436598121  0.323428757  1.292542866
 [31] -0.660441791  0.202075589  1.534192656  0.046520556  0.005255055
 [36]  0.689554560 -0.692102672 -1.408840647  0.330191346 -0.627462016
 [41] -1.379976224 -0.377971300 -1.260394001 -2.254093844 -1.298649241
 [46]  0.699703146  1.216902013  1.006491009 -0.609536210  0.359658813
 [51] -0.847547985  0.083352813 -1.452617420  1.076331469 -1.494341720
 [56] -1.030852673  0.688186204 -0.482923656 -0.188247016  0.716425682
 [61] -0.823105701 -0.340288907 -2.433449626 -0.726518931  0.403210564
 [66] -0.133586945 -0.003357632  1.165596562  0.488298980 -0.835027881
 [71]  0.371354257 -1.064706633  1.338550057 -0.501076507  2.001808024
 [76] -1.009478420 -0.277351796  0.363299071  0.709321360  0.922604386
 [81] -0.268822506  0.918293103 -0.099105591  0.917027670  1.299461068
 [86] -0.700143309 -0.994556925  0.104957197  0.561014156  0.545634635
 [91]  1.886013659 -1.055261045 -1.266781975  0.875876286  0.441661417
 [96] -0.496594242 -0.407019107  1.092515982 -0.550448424 -0.072847493
```

Similar, se pot utiliza cele 4 funcții pentru o distribuție normală generalizată. În argumentul funcțiilor, se precizează valorile pentru cei doi parametri: media și abaterea standard. De exemplu, pentru nota studenților la un examen, se consideră că variabila urmează legea normală cu media egală cu 7 și abaterea standard egală cu 2. Se pot calcula: probabilitatea ca un student să ia nota 5 la examen; probabilitatea ca un student să nu ia examenul; nota până la care se găsesc 90% din studenți. Rezultatele sunt prezentate mai jos.

```
> dnorm(5,7,2)
[1] 0,1209854
> pnorm(5,7,2)
[1] 0,1586553
> qnorm(0.9,7,2)
[1] 9,563103
```

2.8. Teste și exerciții

La finalul acestui capitol, propunem spre rezolvare un set de teste grilă și exerciții.

Teste grilă

1. Conform definiției lui Bernoulli, probabilitatea este:
a) un număr pozitiv cu valori în intervalul [0,1];

b) un raport între numărul de cazuri favorabile şi numărul de cazuri posibile realizării unui eveniment aleator;
c) o convenţie privind şansa de apariţie a unui eveniment.

2. În varianta lui Bernoulli, Legea numerelor mari susţine că:
a) probabilitatea este o limită a frecvenţei relative de apariţie a unui eveniment;
b) probabilitatea este egală cu frecvenţa de apariţie a unui eveniment;
c) în cazul aruncării unui zar de un număr foarte mare de ori, frecvenţa de apariţie a feţei cu numărul 6 se apropie de valoarea 1/6.

3. Pentru două evenimente independente are loc:
a) $P(A \cap B) = P(A) \cdot P(B)$;
b) $P(A \cup B) = P(A) + P(B)$;
c) $P(A \cup B) = P(A) \cdot P(B)$.

4. Se consideră două evenimente date prin submulţimile: A={1,2,4,6}; B={1,2,3,5,6}. Au loc rezultatele:
a) $(A \cap B) = \{1,2,6\}$;
b) $(A \cup B) = \{1,2,3,4,5,6\}$;
c) $(A \cap B) = \{4,5\}$.

5. Se consideră experienţa aleatoare a aruncării a două zaruri. Sunt valabile enunţurile:
a) probabilitatea să apară un punctaj egal cu 12 este 1/36;
b) evenimentele elementare sunt dependente;
c) probabilitatea să apară cuplul şase-şase este egală cu cea a apariţiei cuplului doi-doi;
d) probabilitatea să apară un punctaj egal cu 7 este 1/36.

6. Se consideră două evenimente construite pe baza experienţei aruncării unui zar: A={2,4,6}; B={5,6}. Au loc rezultatele:
a) $P(A \cap B) = 1/6$;
b) $P(A \cup B) = 1/2$;
c) $P(A \cap B) = 1$;
d) $P(A \cup B) = 1$.

BCU "GHEORGHE ASACHI" UNIVERSITY LIBRARY

7. Pentru o variabilă binomială $X \sim B(100;0,5)$, au loc:

a) $E(X)=50$;

b) $E(X)=100$;

c) $V(X)=25$.

8. Pentru o variabilă aleatoare normală standard, au loc:

a) $E(X)=0$;

b) $P(X<-1)=0,159$;

c) $V(X)=1$;

d) $P(X<-1)=0,341$.

Exerciții

1. În cazul experienței aleatoare a aruncării unui zar, se cere să se:

a) scrie legea de distribuție a variabilei care înregistrează numărul de puncte de pe fiecare față a zarului;

b) calculeze probabilitatea să apară o față mai mare decât 2;

c) determine funcția de repartiție pentru distribuția construită.

2. În cazul experienței aleatoare a aruncării a două zaruri, se cere să se:

a) construiască legea de distribuție a variabilei care înregistrează numărul de puncte posibile la orice aruncare;

b) calculeze probabilitatea ca la 10 aruncări să apară numărul de puncte egal cu 12 de 2 ori.

3. Se cunosc datele cu privire la un eșantion de persoane repartizate după gen și consumul de alcool, conform tabelului de mai jos.

	Consumă	Nu consumă	Total
Masculin	42	10	52
Feminin	30	18	48
Total	72	28	100

Să se calculeze:

a) probabilitatea ca o persoană aleasă aleator să fie de gen feminin și consumatoare de alcool;

b) probabilitatea ca o persoană din eșantion să fie consumatoare de alcool;

c) să se completeze întreg tabloul cu probabilități marginale și condiționate pentru datele de mai sus.

BCU "LUCIAN BLAGA" UNIVERSITY LIBRARY

4. În cazul experienței aleatoare a extragerii unei cărți de joc dintr-un pachet, definiți variante de evenimente care să constituie un sistem complet de evenimente. Motivați condiția de exhaustivitate pentru fiecare caz.

5. Pentru o distribuție normală generalizată $X \sim N(4,4)$:

a) să se calculeze $P(X<2)$;

b) să se construiască variabila normală standard corespunzătoare.

6. Utilizând R, pentru o distribuție normală generalizată $X \sim N(4,9)$, să se calculeze:

a) $P(X=6)$;

b) $P(X<5)$;

c) valoarea variabilei până la care se găsesc 25% din valori.

7. Cu ajutorul programului R, generați un eșantion de 400 de valori ale unei variabile aleatoare care urmează o lege de repartiție normală de medie egală cu 4 și dispersie egală cu 9. Reprezentați grafic variabila astfel obținută. Standardizați variabila și construiți o nouă reprezentare grafică. Comparați cele două diagrame.

BCU IASI/CENTRAL UNIVERSITY LIBRARY

		Nu consumă	Total
Masculin	42	10	52
Feminin	30	18	48
Total	72	28	100

BCU IASI/CENTRAL UNIVERSITY LIBRARY

Capitolul 3

REGRESIA LINIARĂ SIMPLĂ

Analiza de regresie vizează construirea modelelor econometrice, numite şi modele de regresie. Acest instrument metodologic fundamental al econometriei s-a dezvoltat mereu, atât în funcţie de datele utilizate, cât şi de problemele analizate.

Acest capitol este dedicat analizei celui mai simplu model econometric: modelul de regresie cu două variabile între care există o legătură de tip liniar sau o dependenţă proporţională. După câteva noţiuni introductive, de clarificare a termenilor şi a naturii acestui tip de analiză, se vor prezenta principalele probleme ale modelării econometrice cu ajutorul modelului de regresie liniară simplă: estimarea parametrilor modelului, testarea modelului de regresie, estimarea şi testarea coeficienţilor de corelaţie.

3.1. Noţiuni introductive

În acest subcapitol, se realizează o scurtă prezentare a istoricului metodei, a tipurilor de regresie şi a naturii metodei de regresie.

3.1.1. Istoric

Origini

Aşa cum îl întâlnim în dicţionare, termenul *regresie* nu sugerează prea multe repere metodologice (conform DEX, regresia este un tip de mişcare înapoi, un pas spre un stadiu sau etapă anterioară a unui proces). În literatura de specialitate, termenul *regresie*, introdus[1] de Francis Galton (1886), admite la început un asemenea sens. Într-un studiu realizat asupra înălţimii populaţiei, acesta arată că, în

[1] F. Galton, „Family Likeness in Stature", *Proceedings of Royal Society*, London, vol. 40, 1886, pp. 42-72.

BCU CLUJ-NAPOCA UNIVERSITY LIBRARY

medie, din părinți de statură înaltă se nasc copii cu o statură mai joasă decât a părinților, adică se realizează o „regresie" a înălțimii spre medie. Formalizarea unui tip de relație între două fenomene sau două variabile reprezintă mecanismul asupra căruia s-a focalizat gândirea lui Galton, iar termenul utilizat de acesta a intrat în istorie, chiar dacă sensul său primar nu mai corespunde cu cel din metodologia științifică actuală.

Sensul modern

În literatura de specialitate, *analiza de regresie* reprezintă o metodă statistică de analiză a dependenței dintre o variabilă, numită *variabilă dependentă*, și una sau mai multe variabile, numite *variabile independente*. În esență, analiza de regresie presupune estimarea unui model care explică modul în care variabila dependentă este influențată de variabila independentă.

În econometrie, pe lângă analiza de regresie este utilizată și *analiza de corelație*, care se referă la studiul intensității legăturii dintre două sau mai multe variabile. Practic, se calculează anumite tipuri de coeficienți de corelație care corespund unor obiective teoretice și practice concrete. În acest capitol, vor fi prezentați câțiva dintre indicatorii de corelație mai frecvent utilizați în analiza statistică și econometrică.

3.1.2. Tipuri de regresie

Legăturile dintre variabilele statistice pot fi clasificate în mai multe categorii, după următoarele criterii: momentul la care se referă, tipul de dependență dintre variabile, numărul variabilelor, tipul (forma) legăturii etc.

Modele de moment și dinamice

Modelul de moment, numit și *model static*, este modelul econometric în care legătura dintre variabile se referă la același moment sau la aceeași perioadă de timp. Pentru construirea acestor modele, se utilizează date din anchete de moment, cum ar fi sondajele statistice, recensămintele sau alte cercetări de moment.

Modelul dinamic este modelul econometric construit pe baza seriilor de timp. Factorul timp apare în model prin precizarea momentelor sau a intervalelor de timp la care se referă datele. Există și modele în care timpul apare ca o variabilă independentă, exprimând trendul seriei de timp.

Modele deterministe și stocastice

Dependența dintre variabile poate fi:

- *deterministă sau funcțională (matematică)*. Asemenea modele sunt mai rar întâlnite, pentru că presupun că între variabile există o legătură de tipul $y_i = f(x_i)$, adică variabila dependentă este explicată în totalitate de

variabilele independente din model. Modelele funcţionale sunt întâlnite în domeniul ştiinţelor naturii, pe când în ştiinţele sociale, se utilizează mai frecvent modelele probabiliste.

- *stocastică sau probabilistă.* În aceste modele, pentru o valoare a variabilei independente, există mai multe valori ale variabilei dependente, determinate probabilistic. În modelele stocastice, variabila dependentă este influenţată şi de o serie de factori care nu apar explicit în model, dar sunt sintetizaţi printr-o variabilă aleatoare numită *variabilă reziduală*. Modelul stocastic este de forma:

$$y_i = f(x_i) + \varepsilon_i .$$

Modele simple şi multiple

Dacă în modelul de regresie apare o singură variabilă independentă, regresia se numeşte *simplă*. Un exemplu de model simplu este modelul care exprimă dependenţa consumului de preţ:

$$C = f(P).$$

Aceste modele sunt întâlnite mai rar în economie, deoarece un fenomen depinde, de regulă, de mai mulţi factori de influenţă. Dacă se alege totuşi un factor determinant, ceilalţi factori pot fi consideraţi implicit prin variabila reziduală.

Dacă în model apar cel puţin două variabile independente, regresia se numeşte *multiplă*. Modelul are forma:

$$Y = f(X_1, X_2) + \varepsilon ,$$

şi este un model multiplu pentru că variabila dependentă este explicată prin influenţa cumulată a celor doi factori care apar în model.

Modele liniare şi neliniare

Modelul liniar este modelul în care relaţia dintre variabile este una de proporţionalitate, legătura dintre variabile fiind descrisă de o funcţie liniară. Exemple de funcţii liniare utilizate în econometrie:

$$Y = \beta_0 + \beta_1 X ,$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 .$$

Modelul neliniar este acela în care legătura dintre variabile este explicată de o funcţie matematică neliniară. Exemple:

$$Y = \beta_0 + \beta_1 \ln X ,$$

$$\ln Y = \beta_0 + \beta_1 X ,$$

$$Y = \beta_0 \cdot X^{\beta_1} ,$$

$$Y = \beta_0 + \beta_1 X^{-1} .$$

BCU IASI / CENTRAL UNIVERSITY LIBRARY

3.1.3. Interpretarea geometrică și statistică a regresiei

Pentru a prezenta semnificația metodei analizei de regresie, se va utiliza exemplul de mai jos.

Exemplu

Considerăm repartiția unei populații de 50 de firme după profitul realizat (variabila dependentă *Y*, exprimată în sute de milioane de lei) și cheltuielile cu publicitatea (variabila independentă X, exprimată în milioane de lei) realizate într-o anumită perioadă de timp (tabelul 3.1).

Tabelul 3.1. *Repartiția firmelor după profit și cheltuielile cu publicitatea*

x_i \ y_j	0-2	2-4	4-6	6-8	Total
60	2	2	-	-	4
70	2	3	-	-	5
80	1	11	7	5	24
90	-	4	5	3	12
100	-	-	3	2	5
Total	**5**	**20**	**15**	**10**	**50**

Sursa: Date convenționale

Pe baza datelor din tabelul 3.1, se pot construi următoarele cinci repartiții condiționate, conform tabelelor de mai jos:

$Y/_{X=60}$

y_j	$y_j^{'}$	n_{ij}
0-2	1	2
2-4	3	2
4-6	5	-
6-8	7	-
Total	-	**4**

$Y/_{X=70}$

$y_j^{'}$	n_{ij}
1	2
3	3
5	-
7	-
Total	**5**

$Y/_{X=80}$

$y_j^{'}$	n_{ij}
1	1
3	11
5	7
7	5
Total	**24**

$Y/_{X=90}$

$y_j^{'}$	n_{ij}
1	-
3	4
5	5
7	3
Total	**12**

$Y/_{X=100}$

$y_j^{'}$	n_{ij}
1	-
3	-
5	3
7	2
Total	**5**

unde $y_j^{'}$ reprezintă mijlocul intervalului $(y_{j-1} - y_j)$.

BCU IASI/CENTRAL UNIVERSITY LIBRARY

Mediile condiţionate ale acestor repartiţii sunt prezentate în următorul tabel:

x_i	$\bar{y}_i$
60	2
70	2,2
80	4,33
90	4,83
100	5,8
Total	-

a. Interpretarea geometrică

Locul geometric al mediilor condiţionate ale variabilei dependente, pentru valori fixate ale variabilei independente, reprezintă o linie poligonală sau o curbă - *linia de regresie*, pentru caz discret, sau *curba de regresie*, pentru caz continuu.

Analiza dependenţei legăturii dintre cele două variabile se poate realiza pe baza unei judecăţi statistice elementare: tipul dependenţei dintre cele două variabile sau modul în care variabila independentă o influenţează pe cea dependentă este sugerat de forma *liniei de regresie statistică* (empirică), construită pe baza mediilor condiţionate, calculate cu ajutorul datelor disponibile.

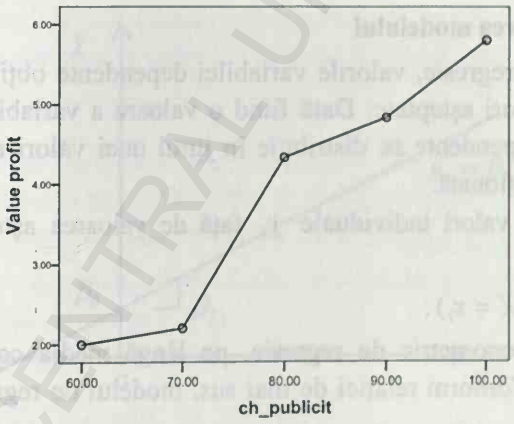


Figura 3.1. *Linia de regresie statistică a lui Y în raport cu X*

Sursa: Realizată de autor (date convenţionale)

Pentru exemplu considerat, se observă că linia de regresie statistică se apropie de o dreaptă, ceea ce sugerează un tip de dependenţă liniară între variabilele studiate (vezi figura 3.1).

BCU "SCIENTIA" UNIVERSITY LIBRARY

b. Interpretarea statistică

Conform teoriei probabilităţilor şi statisticii matematice, regresia este o medie condiţionată definită pe o distribuţie bi- sau multidimensională. În cazul unei legături dintre două variabile, regresia este definită prin aplicaţia:

$E(Y/X=x_i)=f(x_i)$ sau

$E(Y/X)=f(x)$.

Pentru cazul liniar, regresia sau media condiţionată este o funcţie liniară:

$E(Y/X=x_i)=\beta_0+\beta_1 x_i$ sau

$E(Y/X)=\beta_0+\beta_1 X$,

unde β_0, β_1 sunt parametrii modelului, iar X este variabila independentă, considerată o variabilă nestocastică.

3.2. Modelul econometric de regresie liniară simplă

În acest subcapitol, sunt prezentate principalele probleme ale modelării econometrice pe baza celui mai simplu model: modelul cu o singură variabilă independentă aflată într-un raport de dependenţă liniară stocastică cu variabila rezultativă.

3.2.1. Prezentarea modelului

În modelul de regresie, valorile variabilei dependente obţinute prin mediile condiţionate sunt valori aşteptate. Dată fiind o valoare a variabilei independente, valorile variabilei dependente se distribuie în jurul unei valori aşteptate, sperate, care este media condiţionată.

Abaterea unei valori individuale y_i faţă de valoarea aşteptată se notează astfel:

$\varepsilon_i=y_i-E(Y/X=x_i)$.

În modelul econometric de regresie, pe lângă media condiţionată, apare variabila reziduală. Conform relaţiei de mai sus, modelul de regresie econometric este:

$y_i=E(Y/X=x_i)+\varepsilon_i$ sau $Y=E(Y/X)+\varepsilon$.

În cazul regresiei liniare simple, *modelul* are următoarea expresie:

$y_i=\beta_0+\beta_1 x_i+\varepsilon_i$,

dacă este scris pentru fiecare valoare, sau

$Y=\beta_0+\beta_1 X+\varepsilon$,

dacă este scris în general, pentru toate valorile variabilelor.

Media condiţionată are expresia:

BCU IASI / CENTRAL UNIVERSITY LIBRARY

$E(Y/X=x_i)=\beta_0+\beta_1 x_i$ sau

$E(Y/X)=\beta_0+\beta_1 X$.

Pentru a înțelege mai bine natura modelului econometric, se analizează cele două componente ale acestuia, după care se subliniază semnificația parametrilor.

1. Componentele modelului

Modelul econometric liniar simplu include două componente: una deterministă şi una stocastică.

a. Componenta deterministă

Este reprezentată de media condiţionată:

$E(Y/X=x_i)=\beta_0+\beta_1 x_i$.

Notăm regresia liniară simplă astfel (vezi figura 3.2):

$Y_X=\beta_0+\beta_1 X$.

Regresia este reprezentată de o funcție liniară în care apar variabila independentă și parametrii. Aceasta este *regresia la nivelul populației*, fiind determinată de cei doi parametri. Variabila independentă este o variabilă observabilă din punct de vedere statistic, adică nu este o variabilă aleatoare, iar parametrii modelului sunt constante reale.

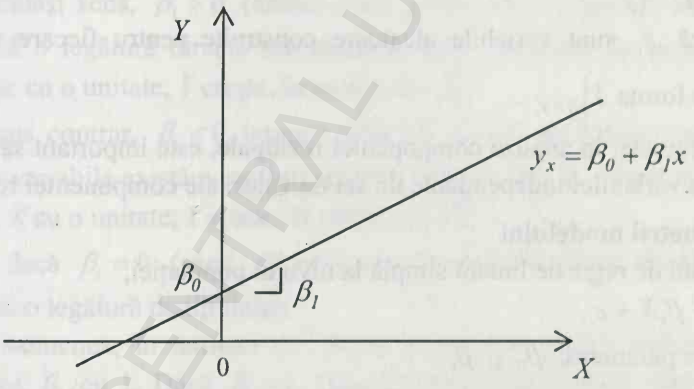


Figura 3.2. *Linia de regresie sau media condiţionată*

b. Componenta aleatoare

În model, aceasta este reprezentată de variabila aleatoare numită *eroare* sau *reziduu* şi este notată cu ε. Natura acestei variabile este legată de următoarele probleme care însoţesc procesul de modelare: natura fenomenului studiat, specificarea modelului, erorile de măsurare[2].

[2] G.S. Maddala, *Introduction to Econometrics*, John Wiley & Sons, 2001, p. 64.

BCU CENTRAL UNIVERSITY LIBRARY

Natura fenomenului studiat

Fenomenele economice au la bază comportamentul uman și influențe care nu sunt ușor de prevăzut. Aceste răspunsuri ale comportamentului uman care nu pot fi prevăzute sunt surprinse prin variabila aleatoare reziduală.

Specificarea modelului

În modelul econometric nu pot apărea toate variabilele sau toți factorii care determină fenomenul măsurat prin variabila dependentă. Există mai multe motive: nu se dispune de date statistice pentru toate variabilele economice, alte variabile sunt imposibil de utilizat, iar altele nu sunt surprinse de cercetători. Prin variabila reziduală, se iau în considerare toate influențele care nu apar explicit în model.

Erorile de măsurare

Variabila reziduală include erorile de măsurare asupra variabilei dependente. Acestea includ erorile cauzate de metoda statistică, în toate etapele demersului care conduc la obținerea datelor necesare modelării.

Observație

Deoarece regresia este o medie condiționată, iar valorile variabilei dependente sunt obținute prin model după relația:

$$y_i = E(Y\big|_{X=x_i}) + \varepsilon_i,$$

rezultă că ε_i sunt variabile aleatoare construite pentru fiecare repartiție condiționată de forma $Y\big|_{X=x_i}$.

Cu alte cuvinte, în analiza componentei reziduale, este important să asociem fiecărei valori a variabilei independente un set de valori ale componentei reziduale.

2. Parametrii modelului

În modelul de regresie liniară simplă la nivelul populației,

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

există doi parametri: β_0 și β_1.

Aceștia se mai numesc și *coeficienți de regresie*. Regula de bază în construirea modelelor este aceea că un parametru este asociat unei variabile.

Parametrul β_0 este asociat unei constante (care ia valoarea egală cu 1). Acest parametru este numit și termenul liber (*intercept*) și indică valoarea medie a variabilei dependente Y, atunci când variabila independentă X ia valoarea zero. Cu alte cuvinte:

$$\beta_0 = E(Y)\big|_{X=0}.$$

Din punct de vedere grafic (vezi figura 3.2), β_0 este ordonata la origine a dreptei de regresie sau intersecția dreptei cu axa *OY*.

Parametrul β_1 (*slope*) indică variația absolută medie a variabilei dependente la o variație absolută cu o unitate a variabilei independente. Cu alte cuvinte, β_1 arată răspunsul (modificarea) variabilei *Y* la o creștere sau scădere cu o unitate a variabilei *X*.

Din punct de vedere geometric, β_1 reprezintă panta dreptei de regresie și se măsoară ca tangentă a unghiului dintre dreapta de regresie și axa *OX* (vezi figura 3.2):

$$\beta_1 = \frac{dY}{dX} \cong \frac{\Delta Y}{\Delta X} \text{ sau } \beta_1 = dY\big|_{dX=1},$$

unde $Y = \beta_0 + \beta_1 X + \varepsilon$.

În alți termeni, conform relației de mai sus, parametrul β_1 reprezintă *viteza de variație* a variabilei dependente în raport cu variabila independentă. Această viteză este constantă în cazul modelului de regresie liniară simplă și exprimă natura dependenței liniare dintre două variabile.

Răspunsul variabilei dependente la variația lui *X* poate fi:

- în același sens, $\beta_1 > 0$ (atunci când dY și dX au același semn), ceea ce indică o legătură directă sau pozitivă între variabile (de exemplu, dacă *X* crește cu o unitate, *Y* crește, în medie, cu β_1);
- în sens contrar, $\beta_1 < 0$ (atunci când dY și dX au semne contrare), adică între variabile există o legătură inversă (caz în care, de exemplu, la o creștere a lui *X* cu o unitate, *Y* scade, în medie, cu β_1);
- nul, dacă $\beta_1 = 0$ (când $dY = 0$), caz în care între cele două variabile nu există o legătură de tip liniar.

De asemenea, în termeni de viteză de variație, putem compara valoarea parametrului β_1 cu 1. Dacă $\beta_1 > 1$, atunci *Y* crește mai repede decât *X*, iar dacă $\beta_1 < 1$, variabila dependentă nu crește la fel de repede ca cea independentă. Dacă $\beta_1 = 1$, atunci la o creștere cu o unitate a variabilei independente, există un efect similar, adică de creștere cu o unitate a variabilei dependente. Comparația pantei dreptei cu 1 are mai mult sens dacă variabilele se exprimă în aceeași unitate de măsură. Se poate ajunge ușor la această situație dacă cele două variabile se standardizează.

Panta dreptei mai poate fi considerată un *factor de proporționalitate*:

$\beta_1 = \frac{dY}{dX}$ sau $dY = \beta_1 \cdot dX$.

Din relația de mai sus, se poate observa că variația absolută a variabilei dependente este proporțională cu cea a variabilei independente. Factorul de proporționalitate este constant și este tocmai panta dreptei. În această relație, se regăsește, de fapt, specificul dependenței liniare dintre două variabile: factorul de proporționalitate rămâne constant, indiferent de variația variabilei independente.

3.2.2. Ipotezele clasice ale modelului de regresie

Modelarea econometrică implică anumite condiții sau ipoteze asupra celor două componente ale modelului, ipoteze care vor fi prezentate în continuare.

Considerăm modelul de regresie liniară simplă:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Ipotezele acestui model se împart în două categorii și privesc cele două componente ale modelului: componenta deterministă și componenta aleatoare.

Ipoteze cu privire la variabila independentă

- variabila independentă X este observabilă, statistică (nestocastică);
- variabila independentă are o dispersie finită și este posibil de determinat.

Ipoteze cu privire la variabila aleatoare eroare

- *eroarea medie este nulă*: $E(\varepsilon_i) = 0$

Cu alte cuvinte, această ipoteză susține că, în medie, modelul este bine specificat, adică factorii neincluși explicit în model nu afectează sistematic valoarea medie a variabilei dependente. Altfel spus, această ipoteză presupune că valoarea așteptată, sperată, a erorii la nivelul repartițiilor condiționate de tipul $Y|X = x_i$ este egală cu zero (vezi figura 3.3).

- *ipoteza de homoscedasticitate*: $V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$

Această ipoteză presupune că varianța erorii este constantă la nivelul repartițiilor condiționate de tipul $Y|_{X=x_i}$. Repartițiile variabilei reziduale pentru fiecare repartiție condiționată sunt prezentate în figura 3.3.

- *ipoteza de normalitate a erorilor*: $\varepsilon_i \sim N(0, \sigma^2)$

La nivelul fiecărei repartiții condiționate, variabila eroare urmează o lege de repartiție normală.

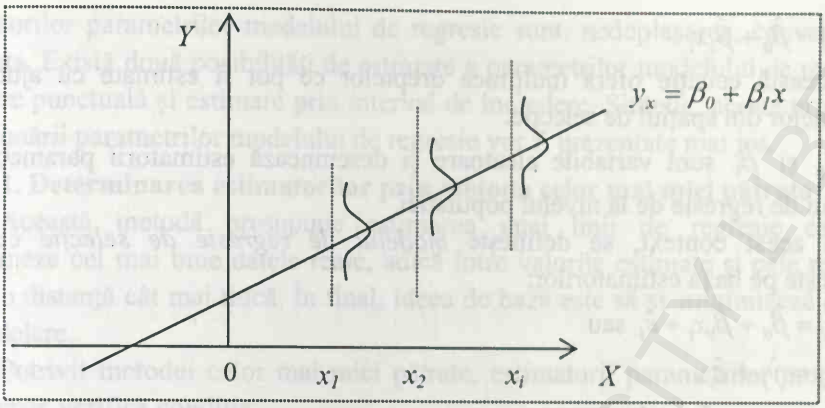


Figura 3.3. *Repartiţiile erorilor la nivelul distribuţiilor condiţionate*

Observaţie

Dacă se admite ipoteza $\varepsilon_i \sim N(0,\sigma^2)$, atunci variabila dependentă este o variabilă aleatoare normal distribuită de forma: $Y \sim N(\beta_0 + \beta_1 X; \sigma^2)$.

- *ipoteza de necorelare a erorilor*: $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$

Ipoteza afirmă că erorile nu se influenţează reciproc, adică sunt independente.

- *lipsa corelaţiei dintre variabila independentă şi variabila eroare*: $\text{cov}(\varepsilon_i, x_i) = 0$.

Analiza acestor ipoteze, în sensul evaluării implicaţiilor nerespectării acestor condiţii asupra calităţii modelului, împreună cu prezentarea metodelor specifice de testare, se va realiza într-un capitol separat al acestei lucrări (vezi capitolul 7).

3.3. Estimarea parametrilor modelului

În practică, de obicei, nu se dispune de date decât de la nivelul unui eşantion de volum *n*. Pe baza acestor date, se realizează estimarea parametrilor modelului de regresie.

Considerăm modelul de regresie liniară simplă:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Regresia la nivelul populaţiei este de forma:

$E(Y / X = x_i) = \beta_0 + \beta_1 x_i$.

Extragerea unui eşantion aleator presupune o mulţime de posibilităţi care este numită spaţiu de selecţie. Pentru mulţimea eşantioanelor posibil de extras, se poate scrie ecuaţia regresiei cu ajutorul estimatorilor parametrilor:

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Această ecuație oferă mulțimea dreptelor ce pot fi estimate cu ajutorul eșantioanelor din spațiul de selecție.

$\hat{\beta}_0$ și $\hat{\beta}_1$ sunt variabile aleatoare și desemnează estimatorii parametrilor modelului de regresie de la nivelul populației.

În acest context, se definește *modelul de regresie de selecție* ce se construiește pe baza estimatorilor:

$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$ sau

$y_i = \hat{y}_i + \hat{\varepsilon}_i$,

unde $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ estimează media condiționată $E(Y/X)$.

Din relațiile de mai sus, rezultă:

$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ sau

$\hat{\varepsilon}_i = y_i - \hat{y}_i$.

$\hat{\varepsilon}_i$ sunt variabile aleatoare care surprind diferențele dintre valorile estimate prin modelele de regresie posibile și valorile reale ale variabilei dependente. Cu alte cuvinte, dacă se dispune de un set de date statistice obținute prin sondaj, se pot calcula erorile estimate ale unui anumit model de regresie, ca diferențe dintre valorile empirice și cele estimate cu ajutorul modelului pentru variabila dependentă.

În concluzie, unui model de regresie la nivelul populației, $Y = \beta_0 + \beta_1 X + \varepsilon$, care depinde de cei doi parametri sau coeficienți de regresie, îi asociem un model de regresie de selecție, $Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\varepsilon}$, care este construit cu ajutorul estimatorilor parametrilor și ai componentei reziduale. În final, la nivelul unui set de date, se obține un model estimat, cu ajutorul estimațiilor parametrilor.

În mod logic, următorul pas metodologic îl reprezintă identificarea estimatorilor care permit estimarea parametrilor modelului și a erorilor de modelare în condiții de calitate bine determinate.

3.3.1. Probleme ale estimării parametrilor modelului de regresie

Problemele pe care le ridică estimarea parametrilor modelului de regresie vizează metodele de determinare a estimatorilor, proprietățile estimatorilor obținuți printr-o anumită metodă și metodele de estimare a parametrilor.

Dintre principalele metode de determinare a estimatorilor parametrilor modelului de regresie, amintim: metoda celor mai mici pătrate, metoda momentelor, metoda verosimilității maxime etc. Principalele proprietăți ale

BCU CENTRAL UNIVERSITY LIBRARY

estimatorilor parametrilor modelului de regresie sunt: nedeplasarea, convergența, eficiența. Există două posibilități de estimare a parametrilor modelului de regresie: estimare punctuală și estimare prin interval de încredere. Sintetic, aceste probleme ale estimării parametrilor modelului de regresie vor fi prezentate mai jos.

1. Determinarea estimatorilor prin metoda celor mai mici pătrate

Această metodă presupune estimarea unei linii de regresie care să aproximeze cel mai bine datele reale, adică între valorile estimate și cele reale să existe o distanță cât mai mică. În final, ideea de bază este să se minimizeze erorile de modelare.

Potrivit metodei celor mai mici pătrate, estimatorii parametrilor modelului de regresie verifică condiția:

$$\sum_i \hat{\varepsilon}_i^2 = \min \text{ sau } \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \min .$$

Prin metoda celor mai mici pătrate, estimatorii parametrilor modelului de regresie liniară simplă se obțin rezolvând problema de optim:

$$S = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \min .$$

Soluția se obține prin respectarea a două condiții: de extrem și de minim, pentru aplicația:

$$S = S(\hat{\beta}_0, \hat{\beta}_1).$$

Condiția de extrem presupune ecuațiile:

$$\begin{cases} \dfrac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = 0 \\ \dfrac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = 0 \end{cases} \text{ sau } \begin{cases} 2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0 \\ 2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0 \end{cases}$$

Rezultă:

$$\begin{cases} \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \sum_i x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

sau

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i = \sum_i y_i \\ \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 = \sum_i y_i x_i \end{cases}$$

BCU SIBIU CENTRAL UNIVERSITY LIBRARY

Condiția de minim presupune ca matricea derivatelor parțiale de ordinul doi ale aplicației S să fie pozitiv definită. Matricea este de tipul:

$$\begin{pmatrix} \frac{\partial^2 S(\hat{\beta}_0,\hat{\beta}_1)}{\partial \hat{\beta}_0^{\,2}} & \frac{\partial^2 S(\hat{\beta}_0,\hat{\beta}_1)}{\partial \hat{\beta}_0 \partial \hat{\beta}_1} \\ \frac{\partial^2 S(\hat{\beta}_0,\hat{\beta}_1)}{\partial \hat{\beta}_1 \partial \hat{\beta}_0} & \frac{\partial^2 S(\hat{\beta}_0,\hat{\beta}_1)}{\partial \hat{\beta}_1^{\,2}} \end{pmatrix}$$

Prin rezolvarea problemei de minim, se obține sistemul de ecuații normale:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i = \sum_i y_i \\ \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 = \sum_i x_i y_i \end{cases}$$

Rezolvarea sistemului conduce la următoarele relații ale estimatorilor:

$$\hat{\beta}_0 = \frac{\sum_i y_i \sum_i x_i^2 - \sum_i x_i \sum_i x_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} \text{ sau}$$

$\hat{\beta}_0 = \hat{\bar{y}} - \hat{\beta}_1 \bar{x}$, iar

$$\hat{\beta}_1 = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} \text{ sau}$$

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \hat{\bar{y}})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\hat{\text{cov}}(X,Y)}{V(X)}.$$

2. Proprietățile estimatorilor $\hat{\beta}_0, \hat{\beta}_1$

Pentru a demonstra proprietățile estimatorilor parametrilor modelului de regresie, vom construi alte relații pentru acești estimatori.

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \hat{\bar{y}})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} - \frac{\sum_i \hat{\bar{y}}\,(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$= \frac{\sum_i y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \sum_i y_i w_i$$

BCU IASI CENTRAL UNIVERSITY LIBRARY

Factorul $w_i = \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$ depinde doar de variabila independentă din model, adică este determinist.

În mod analog, se obţine o formulă pentru estimatorul parametrului β_0, pe baza relaţiei:

$$\hat{\beta}_0 = \hat{\bar{y}} - \hat{\beta}_1 \bar{x} .$$

După efectuarea unor calcule elementare, se obţine:

$$\hat{\beta}_0 = \hat{\bar{y}} - \hat{\beta}_1 \bar{x} = \hat{\bar{y}} - \bar{x} \sum_i y_i w_i = \frac{1}{n} \sum_i y_i - \bar{x} \sum_i y_i w_i = \sum_i y_i \left(\frac{1}{n} - \bar{x} w_i \right) .$$

a. Proprietatea de nedeplasare

Proprietatea de nedeplasare a estimatorilor parametrilor modelului de regresie se demonstrează în condiţiile respectării ipotezei că variabila X este nestocastică şi în baza proprietăţii că variabilele aleatoare y_i urmează aceeaşi lege de repartiţie, adică:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) .$$

Considerăm relaţia:

$$\hat{\beta}_1 = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2} = \sum_i \frac{n x_i - \sum_i x_i}{n \sum_i x_i^2 - (\sum_i x_i)^2} y_i .$$

Rezultă că media estimatorului $\hat{\beta}_1$ este:

$$E(\hat{\beta}_1) = \sum_i \frac{n x_i - \sum_i x_i}{n \sum_i x_i^2 - (\sum_i x_i)^2} \cdot E(y_i) , \text{ iar}$$

$$E(y_i) = \beta_0 + \beta_1 x_i .$$

În consecinţă,

$$E(\hat{\beta}_1) = \beta_1 \cdot \frac{n \sum_i x_i^2 - (\sum_i x_i)^2}{n \sum_i x_i^2 - (\sum_i x_i)^2} = \beta_1 ,$$

ceea ce indică un estimator nedeplasat.

Analog, pentru parametrul β_0, se consideră relaţia:

$$\hat{\beta}_0 = \hat{\bar{y}} - \hat{\beta}_1 \bar{x} = \hat{\bar{y}} - \bar{x} \sum_i y_i w_i = \frac{1}{n} \sum_i y_i - \bar{x} \sum_i y_i w_i = \sum_i y_i \left(\frac{1}{n} - \bar{x} w_i \right) .$$

BCU MS CENTRAL UNIVERSITY LIBRARY

$$E(\hat{\beta}_0) = E(\sum_i y_i(\frac{1}{n} - \bar{x}w_i)) = \sum_i (\frac{1}{n} - \bar{x}w_i)E(y_i).$$

Rezultă:

$E(\hat{\beta}_0) = \sum_i (\frac{1}{n} - \bar{x}w_i)(\beta_0 + \beta_1 x_i)$ şi se poate demonstra imediat că

$E(\hat{\beta}_0) = \beta_0$, deoarece:

$\sum_i (\frac{1}{n} - \bar{x}w_i) = 1$ şi $\sum_i (\frac{1}{n} - \bar{x}w_i)x_i = 0$.

În concluzie:

$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1,$

adică estimatorii $\hat{\beta}_0, \hat{\beta}_1$ sunt nedeplasaţi.

b. Proprietatea de normalitate

Dacă admitem ipoteza că $\varepsilon_i \sim N(0, \sigma^2)$, estimatorii $\hat{\beta}_0, \hat{\beta}_1$, care sunt combinaţii liniare de variabile normal distribuite, sunt normal repartizaţi. Parametrii acestor repartiţii vor fi determinaţi mai jos.

Variabilele aleatoare $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$ au aceeaşi lege de repartiţie cu variabila: $Y = \beta_0 + \beta_1 X + \varepsilon$.

Rezultă:

$E(Y) = \beta_0 + \beta_1 X$, iar

$V(Y) = V(\varepsilon) = \sigma^2$.

Pe baza relaţiilor anterioare, obţinem următorii parametri ai repartiţiilor celor doi estimatori, $\hat{\beta}_0, \hat{\beta}_1$:

$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1,$

$$V(\hat{\beta}_1) = V(\sum_i w_i y_i) = \sum_i w_i^2 V(y_i) = \sum_i w_i^2 \sigma^2 = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2},$$

deoarece: $\sum_i w_i = 0$ şi $\sum_i w_i x_i = 1$.

$$V(\hat{\beta}_0) = V\left(\sum_i y_i(\frac{1}{n} - \bar{x}w_i)\right) = \sigma^2 \sum_i \left(\frac{1}{n} - \bar{x}w_i\right)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right).$$

În concluzie, rezultă următoarele repartiţii ale estimatorilor:

BCU ASSOCIATRAL UNIVERSITY LIBRARY

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right) \text{ sau } \hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1}),$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right)\right) \text{ sau } \hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0}).$$

c. Proprietatea de convergenţă

Estimatorii $\hat{\beta}_0, \hat{\beta}_1$ sunt convergenţi, adică pentru un volum al eşantionului suficient de mare, şirurile estimatorilor converg în probabilitate către parametrii β_0, β_1. Au loc relaţiile:

$$\left(\hat{\beta}_0\right)_{n\in N} \xrightarrow{p} \beta_0,$$

$$\left(\hat{\beta}_1\right)_{n\in N} \xrightarrow{p} \beta_1.$$

d. Proprietatea de eficienţă

Estimatorul $\hat{\beta}_1$ este eficient pentru parametrul β_1, adică, dintre toţi estimatorii posibili, $\hat{\beta}_1$ are varianţa cea mai mică.

Observaţia 1

Se poate demonstra că variabila aleatoare $\hat{\varepsilon}_i = y_i - \hat{y}_i$ urmează o lege de repartiţie normală de forma:

$$\hat{\varepsilon}_i \sim N\left(0, \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)\right).$$

De asemenea, se poate arăta că un estimator nedeplasat al dispersiei erorilor este dat prin relaţia:

$$\hat{\sigma}^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n-2} = \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}, \text{ iar}$$

$$E(\hat{\sigma}^2) = E\left(\frac{\sum_i \hat{\varepsilon}_i^2}{n-2}\right) = \sigma^2,$$

BCU IASI CENTRAL UNIVERSITY LIBRARY

deoarece: $E(\hat{\sigma}^2) = E\left(\frac{\sum_i \hat{\varepsilon}_i^2}{n-2}\right) = \frac{\sum_i E(\hat{\varepsilon}_i^2)}{n-2} = \frac{\sum_i V(\hat{\varepsilon}_i)}{n-2}$.

Observația 2

Considerăm relația de descompunere a variației totale a variabilei dependente, în condițiile existenței legăturii liniare cu variabila independentă:

$\sum_i (y_i - \bar{y})^2 = \sum_i (\beta_0 + \beta_1 x_i - \bar{y})^2 + \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$ sau

$V_T = V_E + V_R$.

Vom nota prin:

$\hat{V}_R = \sum_i \hat{\varepsilon}_i^2 = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$, adică estimatorul variației reziduale.

Dezvoltând relația de mai sus, se poate scrie:

$\hat{V}_R = \sum_i (y_i - \hat{\bar{y}})^2 + \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum_i (x_i - \bar{x})(y_i - \hat{\bar{y}})$, unde

$\hat{V}_T = \sum_i (y_i - \hat{\bar{y}})^2$ este estimatorul variației totale.

Rezultă:

$\hat{V}_R = \hat{V}_T + \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum_i (x_i - \bar{x})(y_i - \hat{\bar{y}})$, iar

$\hat{\beta}_1 = \frac{\sum_i (y_i - \hat{\bar{y}})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\hat{\text{cov}}(X,Y)}{V(X)}$, de unde rezultă:

$\hat{V}_R = \hat{V}_T - \hat{\beta}_1 \sum_i (x_i - \bar{x})(y_i - \hat{\bar{y}})$, iar

$\hat{V}_E = \hat{\beta}_1 \sum_i (x_i - \bar{x})(y_i - \hat{\bar{y}})$, care este estimatorul variației explicate.

Obținem rezultatul:

$\hat{V}_T = \hat{V}_E + \hat{V}_R$.

3.3.2. Estimarea punctuală și prin interval de încredere

Din punct de vedere statistic, estimarea unui parametru se poate realiza punctual și prin interval de încredere.

BCU ASCENTRALUNIVERSITY LIBRARY

a. Estimarea punctuală

În baza proprietăţilor de nedeplasare şi convergenţă, parametrii modelului de regresie se estimează punctual, considerând estimaţiile calculate la nivelul unui eşantion reprezentativ extras din populaţia de referinţă, pe baza relaţiilor:

$$b_0 = \frac{\sum_i y_i \sum_i x_i^2 - \sum_i x_i \sum_i x_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} \text{ sau}$$

$$b_0 = \bar{y} - b_1 \bar{x},$$

$$b_1 = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} \text{ şi}$$

$$\bar{x} = \frac{\sum_i x_i}{n}, \quad \bar{y} = \frac{\sum_i y_i}{n},$$

care reprezintă mediile variabilelor X, Y calculate la nivelul eşantionului.

Pe baza estimaţiilor, se poate scrie *ecuaţia estimată a regresiei*:

$$Y_x = b_0 + b_1 X .$$

b. Estimarea prin interval de încredere a parametrilor β_0, β_1

La baza procedeului de estimare prin interval de încredere stau legile normale de repartiţie ale estimatorilor $\hat{\beta}_0, \hat{\beta}_1$. Astfel, dacă se consideră estimatorii standardizaţi, obţinem statisticile:

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1),$$

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \sim N(0,1).$$

Dacă se utilizează estimatorii abaterilor standard ale estimatorilor, obţinem două variabile care urmează legea de repartiţie Student:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t(n-2),$$

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t(n-2).$$

BCU ASE CENTRAL UNIVERSITY LIBRARY

Conform proprietăților repartiției Student, pentru un nivel de încredere $(1-\alpha)$ fixat, intervalul de încredere pentru parametrul β_1 se determină pe baza relației:

$$P\left(\left|\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_{\hat{\beta}_1}}\right|<t_{\alpha/2}\right)=1-\alpha.$$

Rezultă:

$P(\hat{\beta}_1-t_{\alpha/2}\hat{\sigma}_{\hat{\beta}_1}\le\beta_1\le\hat{\beta}_1+t_{\alpha/2}\hat{\sigma}_{\hat{\beta}_1})=1-\alpha$, unde

$$\hat{\sigma}_{\hat{\beta}_1}=\sqrt{\frac{\hat{\sigma}^2}{\sum_i(x_i-\bar{x})^2}},\ \text{iar}$$

$$\hat{\sigma}^2=\frac{\sum_i\hat{\varepsilon}_i^2}{n-2}=\frac{\sum_i(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2}{n-2}\ \text{sau}$$

$$\hat{\sigma}^2=\frac{\hat{V}_R}{n-2}.$$

Cu alte cuvinte, pentru un nivel de încredere egal cu $(1-\alpha)$, limitele intervalului de încredere pentru parametrul β_0 sunt:

$$\hat{\beta}_0\pm t_{\alpha/2}\hat{\sigma}_{\hat{\beta}_0}.$$

Analog, pentru parametrul β_1, intervalul de încredere este:

$$\hat{\beta}_1\pm t_{\alpha/2}\hat{\sigma}_{\hat{\beta}_1}.$$

Pe baza datelor de la nivelul unui eşantion, se calculează un interval de încredere cu ajutorul estimaţiilor. Se obţin intervalele estimate:

$b_0\pm t_{\alpha/2}s_{\hat{\beta}_0}$, respectiv

$b_1\pm t_{\alpha/2}s_{\hat{\beta}_1}$.

Estimaţiile pentru abaterile standard ale estimatorilor sunt:

$$s_{\hat{\beta}_0}=\sqrt{s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum_i(x_i-\bar{x})^2}\right)},$$

$$s_{\hat{\beta}_1}=\sqrt{\frac{\sum_i(y_i-b_0-b_1x_i)^2}{(n-2)\sum_i(x_i-\bar{x})^2}}=\sqrt{\frac{s^2}{\sum_i(x_i-\bar{x})^2}},\ \text{iar}$$

BCU IAŞI CENTRAL UNIVERSITY LIBRARY

$s = \sqrt{\dfrac{\sum_i (y_i - b_0 - b_1 x_i)^2}{(n-2)}}$ este estimaţia parametrului σ.

Dacă notăm cu $e_i = y_i - b_0 - b_1 x_i$ estimaţiile erorilor, estimaţia parametrului σ devine:

$$s = \sqrt{\frac{\sum_i e_i^2}{(n-2)}}.$$

Pentru componentele variaţiei, se obţin următoarele estimaţii:

$TSS = \sum_i (y_i - \bar{y})^2$ (*Total Sum of Squares*);

$ESS = \sum_i (b_0 + b_1 x_i - \bar{y})^2$ (*Explained Sum of Squares*);

$RSS = \sum_i (y_i - b_0 - b_1 x_i)^2 = \sum_i e_i^2$ (*Residual Sum of Squares*).

Pentru aceste estimaţii, are loc relaţia:

$TSS = ESS + RSS$.

3.3.3. Regresia liniară simplă în R

Cu ajutorul programului R, se pot estima modele de regresie simplă, pe baza datelor statistice disponibile pentru variabilele analizate. De asemenea, R permite reprezentarea grafică a distribuţiei bidimensionale şi a liniei de regresie care exprimă legătura dintre variabile. Funcţia care permite construirea unui model de regresie liniară în R este următoarea:

```
> lm(formula, dataset)
```

Primul element din argumentul funcţiei este *formula.* Aceasta presupune prezentarea în ordine a variabilelor din model, conform relaţiei:

```
variabila dependentă ~ variabila independentă.
```

Funcţia adaugă implicit constanta în model, de aceea nu este necesar să se adauge o specificaţie în acest sens. Opţiunea privind setul de date poate fi neglijată dacă în prealabil s-a încărcat în spaţiul de lucru o structură de date bine definită (folosind comanda *attach*).

Exemplul 1

Ca exemplu, considerăm datele prezentate în capitolul 1 cu privire la repartiţia judeţelor României după cele mai importante rate demografice (vezi tabelul 3.2). Din această structură de date ne oprim la două variabile: rata de

B.C.U. M. EMINESCU" IAŞI

natalitate (variabila dependentă *Y*, exprimată în număr de născuți vii la 1000 de locuitori) și rata de nupțialitate (variabila independentă *X*, exprimată în număr de căsătorii la 1000 de locuitori).

Tabelul 3.2. *Repartiția județelor după principalele rate demografice, în 2014*

R R Console

	Judetul	Nat	Mort	Spor	Cas	Div	Inf.
1	Bihor	9.3	11.7	-2.4	5.1	1.20	6.6
2	Bistrita-Nasaud	10.0	10.0	0.0	5.5	0.60	6.7
3	Cluj	9.1	10.9	-1.8	7.1	0.97	7.0
4	Maramures	8.7	10.5	-1.8	5.7	1.32	7.4
5	Satu Mare	9.1	10.8	-1.7	5.4	1.26	7.8
6	Salaj	10.1	11.9	-1.8	5.1	1.10	10.4
7	Alba	8.2	11.3	-3.1	4.9	1.34	8.3
8	Brasov	9.8	9.7	0.1	5.4	1.78	7.1
9	Covasna	9.8	10.9	-1.1	4.4	0.71	7.6
10	Harghita	10.0	10.6	-0.6	4.3	0.55	9.0
11	Mures	9.5	11.4	-1.9	4.5	1.04	10.2
12	Sibiu	9.7	9.9	-0.2	5.7	1.22	8.7
13	Bacau	8.8	11.3	-2.5	4.9	1.34	7.5
14	Botosani	9.0	12.7	-3.7	4.5	1.62	12.8
15	Iasi	10.3	9.5	0.8	5.7	1.16	7.9
16	Neamt	8.4	11.3	-2.9	4.7	1.37	9.5
17	Suceava	10.7	10.4	0.3	6.6	1.30	8.7
18	Vaslui	8.8	11.2	-2.4	4.7	1.40	9.5
19	Braila	6.5	13.1	-6.6	4.0	1.95	10.2
20	Buzau	7.7	13.8	-6.1	4.0	1.09	10.4
21	Constanta	9.5	10.2	-0.7	5.4	1.66	11.6
22	Galati	7.8	10.7	-2.9	5.0	1.18	8.2
23	Tulcea	8.1	12.5	-4.4	4.5	0.83	11.0
24	Vrancea	8.8	12.0	-3.2	4.4	0.91	11.6
25	Arges	8.0	11.2	-3.2	5.1	1.39	8.1
26	Calarasi	8.5	13.7	-5.2	3.9	1.29	16.2
27	Dâmbovita	8.4	11.8	-3.4	4.5	1.03	7.6
28	Giurgiu	8.2	15.2	-7.0	4.4	0.53	13.5
29	Ialomita	9.2	13.2	-4.0	4.0	0.91	9.9
30	Prahova	8.0	12.0	-4.0	4.8	1.33	7.7
31	Teleorman	7.3	17.0	-9.7	3.4	1.13	10.4
32	Ilfov	11.0	10.1	0.9	6.4	1.28	4.7
33	Bucuresti	8.9	10.1	-1.2	7.4	1.31	5.0
34	Dolj	8.3	13.6	-5.3	4.7	0.63	7.9
35	Gorj	7.4	11.6	-4.2	5.4	1.55	9.8
36	Mehedinti	7.9	13.8	-5.9	5.0	0.85	8.3
37	Olt	7.3	14.2	-6.9	4.4	0.93	7.5
38	Vâlcea	7.3	10.1	-2.8	3.5	0.98	9.5
39	Arad	8.5	12.2	-3.7	5.7	1.12	7.9
40	Caras-Severin	7.0	12.3	-5.3	5.2	1.66	9.9
41	Hunedoara	7.1	12.3	-5.2	4.7	1.55	7.5
42	Timis	9.2	10.2	-1.0	6.4	1.10	7.2

Sursa: Realizat pe baza datelor oferite de INS, 2016

Reprezentăm grafic legătura dintre cele două variabile (Figura 3.4) cu ajutorul comenzilor:

```
> plot(Nat~Cas)
> abline(lm(Nat~Cas))
```

Pentru a estima modelul liniar de regresie dintre natalitate și nupțialitate, utilizăm comanda de mai jos ce presupune crearea unui obiect care memorează modelul de regresie dintre cele două variabile:

```
> model1=lm(Nat~Cas)
```

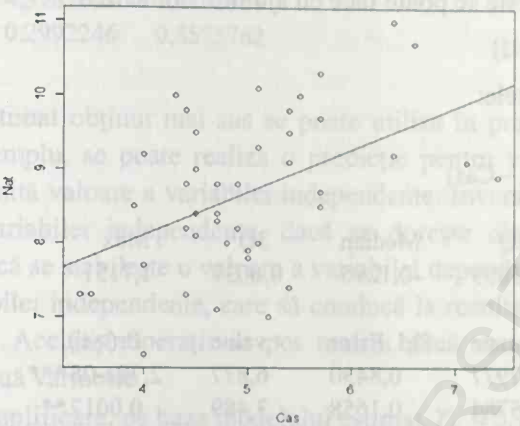


Figura 3.4. *Distribuția județelor după natalitate și nupțialitate*
Sursa: Realizată de autor pe baza datelor oferite de INS, 2016

Estimațiile parametrilor se obțin tastând numele obiectul creat:

```
> model1
```

Rezultatul obținut este prezentat mai jos:

```
Call:
lm(formula = Nat ~ Cas)
Coefficients:
(Intercept)     Cas
   5,7977     0,5784
```

Modelul estimat pentru cele două variabile este de forma:

$Y_X = 5{,}79 + 0{,}57X$.

Valoarea pozitivă a estimației parametrului β_1 indică o legătură directă între natalitate și nupțialitate, așa cum se poate observa și din analiza reprezentării grafice din figura 3.4.

Valorile estimațiilor arată că:

- în cazul unei nupțialități nule ($X = 0$), la nivelul unui județ se înregistrează o rată brută de natalitate medie de 5,79 născuți vii la 1000 de locuitori;
- la o creștere a nupțialității cu o unitate (o căsătorie la 1000 de locuitori), rata brută de natalitate crește, în medie, cu 0,57 născuți vii la 1000 de locuitori.

Valoarea estimată a parametrului β_1 indică viteza de variație a natalității în raport cu nupțialitatea. Aceasta este subunitară, ceea ce înseamnă că natalitatea

BCU CENTRAL UNIVERSITY LIBRARY

crește mai lent decât nupțialitatea. Vizualizarea rezultatelor obținute pentru modelul liniar de regresie se poate face cu ajutorul comenzii:

```
> summary(model1)
```

Se obțin rezultatele:

```
Call:
lm(formula = Nat ~ Cas)
Residuals:
     Min       1Q   Median       3Q      Max
 -1,8054  -0,5895  -0,1295   0,6039   1,7151
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  5,7977    0,8430      6,877    2,78e-08***
Cas          0,5784    0,1658      3,489    0,0012**
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 0,9369 on 40 degrees of freedom
Multiple R-squared: 0,2333,     Adjusted R-squared: 0,2141
F-statistic: 12,17 on 1 and 40 DF,  p-value: 0,001196
```

Pe baza rezultatelor de mai sus, se pot calcula intervalele de încredere ale parametrilor modelului, utilizând estimațiile parametrilor și estimațiile abaterilor standard ale estimatorilor care sunt prezentate în coloana *Std. Error*.

Calculul automat al intervalelor de încredere (cu un nivel de încredere implicit de 0,95) pentru cei doi parametri se realizează utilizând comanda:

```
> confint(model1)
```

Se obțin rezultatele:

```
              2,5%        97,5%
(Intercept)  4,0938637   7,5015915
Cas          0,2433141   0,9134866
```

Intervalele de încredere pentru cei doi parametri au următoarea interpretare, cu un nivel de încredere de 95%:

- parametrul β_0 este acoperit de intervalul *(4,09; 7,50)*;
- parametrul β_1 este acoperit de intervalul *(0,24; 0,91)*.

Dacă se dorește utilizarea unui alt nivel de încredere, acesta se specifică în argumentul funcției. De exemplu, dacă se dorește o probabilitate de 0,9, comanda este următoarea:

```
> confint(model1, level =0.90)
```

Mărimea intervalelor se modifică după cum urmează:

BCU "CAROL I" TRANSILVANIA UNIVERSITY LIBRARY

	5%	95%
(Intercept)	4,3781601	7,2172951
Cas	0,2992246	0,8575762

Observaţie

Modelul estimat obţinut mai sus se poate utiliza în practică pentru diverse obiective. De exemplu, se poate realiza o predicţie pentru variabila dependentă, cunoscând o anumită valoare a variabilei independente. Invers, se poate realiza un control asupra variabilei independente, dacă se doreşte obţinerea unui anumit rezultat: adică, dacă se stabileşte o valoare a variabilei dependente, se poate calcula o valoare a variabilei independente, care să conducă la rezultatul aşteptat, pe baza ecuaţiei estimate. Aceleaşi operaţii se pot realiza dacă se consideră o anumită variaţie a celor două variabile.

Pentru exemplificare, pe baza modelului estimat $Y_X = 5,79 + 0,57X$, se cere:

- să se estimeze rata natalităţii la nivelul unui judeţ pentru care rata nupţialităţii este de 8 căsătorii la 1000 de locuitori;
- să se precizeze cât ar trebui să fie nivelul nupţialităţii pentru a obţine o natalitate de 13 născuţi vii la 1000 de locuitori;
- să se estimeze cu cât creşte nivelul natalităţii dacă are loc o creştere a nupţialităţii cu 1,5 căsătorii la 1000 de locuitori;
- să se determine valoarea cu care ar trebui să crească nivelul nupţialităţii pentru a avea o creştere medie a natalităţii cu 2 născuţi la 1000 de locuitori.

Rezolvare

- dacă $X = 8$, pe baza ecuaţiei estimate a modelului de regresie, se obţine estimaţia cerută: $Y_X = 5,79 + 0,57 \cdot 8 = 10,35$ născuţii vii la 1000 de locuitori;
- pentru o valoare $Y = 13$ născuţi vii la 1000 de locuitori, se calculează: $13 = 5,79 + 0,57X$, de unde rezultă $X = 12,64$ căsătorii la 1000 de locuitori;
- se cunoaşte că $b_1 = 0,57 = dY\big|_{dX=1}$ sau $0,57 = \frac{dY}{dX}$, de unde se obţine $dY = 0,57dX$. Pentru $dX = 1,5$, rezultă $dY = 0,85$ născuţi vii la 1000 de locuitori, care reprezintă creşterea natalităţii corespunzătoare creşterii nupţialităţii cu 1,5 căsătorii la 1000 de locuitori;
- dacă $dY = 2$ şi $0,57 = \frac{dY}{dX}$, rezultă $dX = \frac{dY}{0,57} = \frac{2}{0,57}$, deci $dX = 3,51$ căsătorii la 1000 de locuitori, ceea ce reprezintă creşterea nupţialităţii care ar permite o creştere medie estimată a natalităţii cu 2 născuţi vii la 1000 de locuitori.

Exemplul 2

Considerăm un alt exemplu cu privire la repartiția județelor României, în anul 2014, după rata șomajului (%) și PIB pe locuitor (lei). Datele sunt oferite de www.insse.ro.

Prezentăm codul R și rezultatele obținute pentru estimarea unui model de regresie liniară între cele două variabile.

```
> econ <- read.csv("~/Desktop/economic.csv")
> attach(econ)
> plot(som~pib)
> abline(lm(som~pib))
```

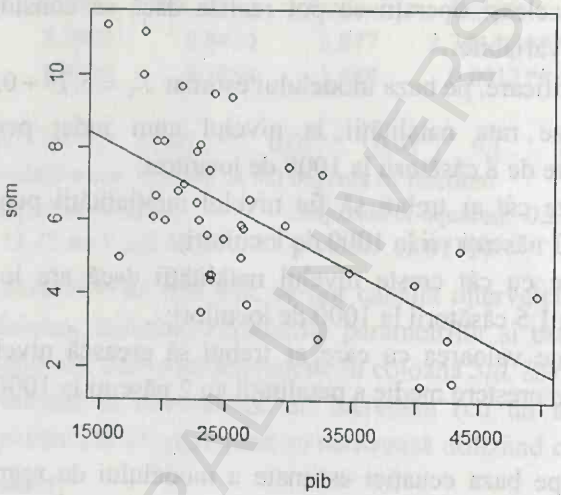


Figura 3.5. *Distribuția județelor după rata șomajului și PIB pe locuitor*
Sursa: Realizată de autor pe baza datelor oferite de INS, 2016

Reprezentarea grafică de mai sus arată că între cele două variabile există o legătură inversă. Pe măsură ce crește nivelul de dezvoltare economică (măsurat prin PIB pe locuitor) pentru un județ, rata șomajului este tot mai mică. Ne așteptăm, așadar, ca estimația pantei dreptei de regresie să fie negativă.

```
> summary(lm(som~pib))

Call:
lm(formula = som ~ pib)
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  9,370e+00   7,664e-01   12,225    4,35e-15***
pib         -1,172e-04   2,509e-05   -4,672    3,35e-05***
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

Ecuația estimată a modelului de regresie este următoarea:
$Y_X = 9{,}37 - 0{,}000117X$.

Rezultatul ne arată că o creștere a PIB pe locuitor, la nivelul unui județ, cu 10000 de lei conduce la o scădere, în medie, a ratei șomajului cu 1,17%.

```
> confint(lm(som~pib))
                2,5%          97,5%
(Intercept) 7,8207155523  1,091876e+01
pib        -0,0001679067 -6,650359e-05
```

3.4. Indicatori de corelație

În analiza econometrică, se utilizează următorii indicatori de corelație mai importanți: coeficie

ntul de corelație Pearson, raportul sau coeficientul de determinație, raportul de corelație.

3.4.1. Coeficientul de corelație

Coeficientul de corelație măsoară intensitatea legăturii dintre cele două variabile. În cazul unei legături liniare dintre două variabile, se definește coeficientul de corelație Pearson prin relația:

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sqrt{V(X)V(Y)}} \text{ sau}$$

$$\rho = \frac{N\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\sqrt{\left(N\sum_i x_i^2 - (\sum_i x_i)^2\right)\left(N\sum_i y_i^2 - (\sum_i y_i)^2\right)}}.$$

În baza relației de definiție, coeficientul ia valori în intervalul [-1,1]:
$-1 \le \rho \le 1$.

Dacă valoarea parametrului se apropie de unu, între variabile există o legătură intensă sau puternică. Legătura este slabă atunci când coeficientul are o valoare apropiată de zero. Se consideră semnificativă intensitatea legăturii dintre două variabile dacă $|\rho| \ge 0{,}7$. Pentru valori în jurul lui 0,5, vorbim despre o legătură de intensitate medie. Semnul coeficientului indică sensul legăturii dintre variabile: o valoare pozitivă indică o legătură directă între variabile, iar o valoare negativă indică o legătură inversă.

BCU "CAROL I" CENTRAL UNIVERSITY LIBRARY

Observație

O altă relație pentru coeficientul de corelație se poate construi ținând cont de relația care are la bază coeficientul de regresie β_1:

$$\rho = \beta_1 \sqrt{\frac{V(X)}{V(Y)}} \text{ sau } \rho = \beta_1 \frac{\sigma_X}{\sigma_Y}.$$

Estimarea coeficientului de corelație

Pentru acest parametru, se poate construi un estimator pe baza relației de mai sus:

$$\hat{\rho} = \hat{\beta}_1 \sqrt{\frac{V(X)}{\hat{V}(Y)}}.$$

O estimație a coeficientului de corelație se obține la nivelul unui eșantion, pe baza relației:

$$r = b_1 \sqrt{\frac{s_x'^2}{s_y'^2}} \text{ sau } r = b_1 \frac{s_x'}{s_y'}.$$

Observație

Dacă se realizează o standardizare a variabilelor X, Y, atunci estimatorul coeficientului de corelație pentru aceste variabile este identic cu cel al coeficientului de regresie β_1. Estimația pantei dreptei de regresie pentru un model cu variabile standardizate este identică cu estimația coeficientului de corelație:

$r = \tilde{b}_1$.

Exemplul 1

Pentru datele prezentate în tabelul 3.2 cu privire la repartiția județelor României după rata de natalitate (variabila dependentă Y, exprimată în număr de născuți vii la 1000 de locuitori) și rata de nupțialitate (variabila independentă X, exprimată în număr de căsătorii la 1000 de locuitori), se poate estima coeficientul de corelație Pearson utilizând R în cel puțin două moduri.

În primul rând, coeficientul de corelație se poate obține utilizând comanda:

```
> cor(Nat, Cas)
```

Rezultă estimația:

```
[1] 0,4829942
```

Valoarea estimată a coeficientului de corelație dintre cele două variabile este egală cu 0,482. Aceasta este pozitivă și indică o legătură directă între natalitate și nupțialitate, de intensitate medie.

În al doilea rând, coeficientul de corelație se poate estima construind modelul de regresie pentru cele două variabile standardizate. În R, se utilizează o comandă care returnează coeficienții de regresie pe baza unei funcții numite *lm.beta* (aceasta este disponibilă în pachetul *QuantPsyc*):

```
> lm.beta(model1)
```

Se obține rezultatul:

```
Cas
0,4829942
```

Prin urmare, estimația coeficientului de regresie pentru modelul liniar construit cu variabile standardizate este egală cu estimația coeficientului de corelație:

$\widetilde{b}_1 = r = 0{,}482$.

Observație

Coeficientul de corelație estimat se poate calcula efectiv pe baza relației:

$$r = b_1 \sqrt{\frac{s_x'^2}{s_y'^2}},$$

dacă se utilizează estimațiile dispersiilor celor două variabile și estimația pantei dreptei de regresie.

În R, cele două varianțe se obțin prin comenzile:

```
> var(Nat)
[1] 1,11705
> var(Cas)
[1] 0,77893
```

Cunoscând valoarea lui b_1, egală cu 0,57, din calcule rezultă $r = 0{,}482$.

Exemplul 2

Pentru datele prezentate la exemplul 2 din secțiunea anterioară, prezentăm rezultatele analizei de corelație împreună cu comenzile din R.

```
> cor(som,pib)
[1] -0,623769
> lm.beta(lm(som~pib))
pib
-0,623769
```

BCU "ALECU RUSSO" UNIVERSITY LIBRARY

După cum se observă, coeficientul de corelație estimat între cele două variabile este $r = -0{,}623$ și indică o legătură inversă de intensitate medie între șomaj și PIB pe locuitor la nivelul județelor României. De asemenea, se observă că acest coeficient este identic cu estimația coeficientului de regresie din modelul cu ambele variabile standardizate:

$\widetilde{b}_1 = r = -0{,}623$.

Ca interpretare, se poate spune că, la o creștere a PIB pe locuitor cu o abatere standard, rata șomajului înregistrează o scădere cu 0,623 abateri standard.

3.4.2. Raportul de determinație și raportul de corelație

Urmează să prezentăm doi indicatori de corelație care se pot determina pentru orice tip de model econometric, nu doar pentru cel liniar. Aceștia au la bază proprietatea de descompunere a variației totale a variabilei dependente în cele două componente: variația explicată și variația reziduală.

Raportul de determinație (coeficientul de determinație)

Raportul de determinație este un parametru care se calculează pe baza valorilor reale (y_i) și a valorilor teoretice ($y_{x_i} = \beta_0 + \beta_1 x_i$), valori calculate cu ajutorul modelului de regresie pentru variabila dependentă.

Raportul de determinație măsoară cât din variația totală a variabilei dependente este explicat de modelul de regresie. Relația acestui indicator este următoarea:

$$\eta^2 = \frac{\sum_i (y_{x_i} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{V_E}{V_T} = 1 - \frac{V_R}{V_T},$$

unde: $0 < \eta^2 < 1$.

Exprimată în procente, valoarea raportului de determinație arată cât la sută din variația variabilei dependente este explicată de variația variabilei independente. Acest indicator se poate determina pentru orice tip de model de regresie, liniar sau neliniar.

Estimarea raportului de determinație

La nivelul unei selecții de volum n, raportul de determinație este estimat pe baza relației de descompunere a estimatorului variației totale:

$$\sum_i (y_i - \hat{\bar{y}})^2 = \sum_i (\hat{y}_i - \hat{\bar{y}})^2 + \sum_i (y_i - \hat{y}_i)^2 \text{ sau}$$

$$\hat{V}_T = \hat{V}_E + \hat{V}_R.$$

BCU BUCUREȘTI CENTRAL UNIVERSITY LIBRARY

Observaţie

Deoarece variabila dependentă urmează o lege de repartiţie normală, de parametri $(\beta_0 + \beta_1 X, \sigma^2)$, pentru estimatorii de mai sus, se poate demonstra că sunt variabile cu legi de repartiţie cunoscute:

$\hat{V}_T \sim \chi^2(n-1)$,

$\hat{V}_E \sim \chi^2(k-1)$,

$\hat{V}_R \sim \chi^2(n-k)$,

unde k este numărul de parametri incluşi în model, iar n este volumul eşantionului. Pentru modelul liniar simplu, $k = 2$.

Estimatorul raportului de determinaţie se defineşte ca raport între estimatorul variaţiei explicate şi estimatorul variaţiei totale. În aceste condiţii, se poate scrie relaţia:

$$\hat{\eta}^2 = \frac{\hat{V}_E}{\hat{V}_T} = 1 - \frac{\hat{V}_R}{\hat{V}_T}.$$

Dacă se ia în considerare relaţia:

$$\hat{V}_E = \hat{\beta}_1 \sum_i (x_i - \bar{x})(y_i - \hat{\bar{y}}),$$

estimatorul coeficientului de determinaţie se mai poate scrie astfel:

$$\hat{\eta}^2 = \hat{\beta}_1 \frac{\sum_i (x_i - \bar{x})(y_i - \hat{\bar{y}})}{\sum_i (y_i - \hat{\bar{y}})^2}.$$

Estimaţia raportului de determinaţie este notată cu R^2 şi are expresia:

$$R^2 = \frac{\sum_i (b_0 + b_1 x_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Observaţie

Pentru modelul liniar simplu, au loc relaţiile:

$\rho^2 = \eta^2$, $r^2 = R^2$ sau

$|r| = R$.

Raportul de corelaţie

Indicatorul $\eta = \sqrt{\eta^2}$ se numeşte *raport de corelaţie* şi măsoară intensitatea legăturii dintre cele două variabile.

BCU "ASCENTRAL UNIVERSITY" LIBRARY

Raportul de corelație respectă condiția: $0 \leq \eta \leq 1$.

Estimația raportului de corelație se notează cu R și se obține ca rădăcină pătrată a estimației coeficientului de determinație:

$R = \sqrt{R^2}$.

Raportul de corelație se interpretează în funcție de mărimea acestuia:

- pentru o valoare apropiată de zero, între variabile există o legătură de intensitate slabă;
- pentru o valoare apropiată de unu, legătura este puternică. Se consideră că legătura dintre variabile este semnificativă dacă $R > 0,7$.

Exemplu

Pentru datele din tabelul 3.2, estimația pentru raportul de determinație se prezintă în setul de rezultate obținut pentru modelul liniar de regresie, cu ajutorul comenzii:

```
> summary(model1)
```

Redăm doar ultima parte a rezultatelor returnate de R:

```
Residual standard error: 0,9369 on 40 degrees of freedom
Multiple R-squared: 0,2333,        Adjusted R-squared: 0,2141
F-statistic: 12,17 on 1 and 40 DF,  p-value: 0,001196
```

Valoarea estimată a raportului de determinație (*R Square*) este egală cu 0,2333 și arată că 23,33% din variația variabilei dependente este explicată de variația variabilei independente inclusă în model. Deoarece legătura dintre variabile este una directă ($\beta_1 > 0$), estimația raportului de corelație este egală cu cea a coeficientului de corelație, adică $r = R = \sqrt{R^2} = 0,482$, ceea ce indică o legătură de intensitate medie între cele două variabile.

Observație

Estimația raportului de determinație se poate calcula și pe baza tabelului de rezultate în care sunt prezentate estimațiile componentelor variației: *TSS*, *ESS* și *RSS*.

În R, rezultatele se obțin cu ajutorul funcției *anova* aplicate obiectului *model1*, adică obiectului rezultat prin funcția ce construiește modelul liniar de regresie. Comanda în R este următoarea:

```
> anova(model1)
```

Rezultatele analizei sunt prezentate în tabelul Anova:

```
Analysis of Variance Table
Response: Nat
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cas	1	10,684	10,6842	12,171	0,001196 ***
Residuals	40	35,115	0,8779		

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

În tabelul Anova, estimaţiile componentelor variaţiei sunt prezentate în coloana *Sum Sq*. Variaţia explicată estimată este *10,68* (*Explained Sum of Squares* sau *Regression Sum of Squares*), iar variaţia reziduală estimată este *35,11* (*Residual Sum of Squares*). Variaţia totală estimată se obţine ca sumă a celor două componente:

$TSS = ESS + RSS = 45,79.$

Pe baza acestor valori, utilizând relaţia $R^2 = \frac{ESS}{TSS}$, se obţine:

$$R^2 = \frac{10,68}{45,79} = 0,233.$$

Dispunând de datele din tabelul Anova, se poate obţine valoarea absolută a coeficientului de corelaţie:

$$|r| = R = \sqrt{R^2} = \sqrt{\frac{ESS}{TSS}} = \sqrt{0,233} = 0,482.$$

3.5. Testarea parametrilor şi a modelului de regresie

Testarea parametrilor modelului de regresie, precum şi a modelului de regresie, se realizează după schema clasică a unui procedeu de testare, ale cărei etape sunt precizate în continuare.

3.5.1. Etapele testării

Etapele procesului de testare a unei ipoteze statistice sunt:

1. formularea ipotezelor (ipoteza nulă şi ipoteza alternativă);
2. alegerea pragului de semnificaţie α sau a limitei erorii de speţa întâi (eroarea de a respinge ipoteza nulă, în condiţiile în care aceasta este adevărată);
3. alegerea statisticii test adecvate, care, în condiţiile acceptării ipotezei nule, are o lege de repartiţie specificată;
4. determinarea unei valori teoretice a testului, în funcţie de legea de repartiţie şi de pragul de semnificaţie ales. Pe baza valorii teoretice a testului, se stabilesc regiunile de acceptare şi de respingere a ipotezei nule;
5. calcularea unei valori a statisticii test, pe baza datelor de la nivelul unui eşantion;
6. aplicarea regulii de decizie de acceptare sau de respingere a ipotezei nule (care, în esenţă, constă în compararea valorii calculate a testului cu cea teoretică

sau identificarea poziției acestei valori în una dintre regiunile stabilite în etapa 4).

Regula de decizie cu privire la ipoteza nulă (respingerea sau nu a acesteia) se poate formula în două moduri:

- prin compararea valorii calculate a testului cu valoarea teoretică;
- prin compararea semnificației testului cu pragul de semnificație fixat în etapa a doua a procesului de testare.

Valoarea teoretică se citește pentru un prag de semnificație ales și pentru o statistică cu legea de repartiție cunoscută. De exemplu, pentru legea Student cu n grade de libertate, un prag de semnificație α și un test unilateral, valoarea teoretică ($t_{\alpha,n}$) are proprietatea:

$$P(t > t_{\alpha,n}) = \alpha .$$

Pentru un test bilateral, se consideră proprietatea:

$P(|t| > t_{\alpha/2,n}) = \alpha$ sau $P(t > t_{\alpha/2,n}) = \alpha/2$.

Nivelul de semnificație reprezintă o probabilitate calculată în procesul de testare cu ajutorul tabelelor probabiliste specifice repartiției testului utilizat.

Calculul nivelului de semnificație (p-value sau Sig)

Probabilitatea calculată reprezintă probabilitatea asociată valorii calculate a testului și a primit numele de semnificație a testului. În pachetele program de statistică, aceasta este notată cu *p-value* sau *Sig*.

Pentru o statistică Student, de exemplu, în cazul unui test unilateral, *Sigt* este probabilitatea:

$$Sigt = P(t > t_{calc}),$$

iar aceasta reprezintă probabilitatea cu care nu se respinge ipoteza nulă.

În cazul unul test bilateral, semnificația testului este dată de:

$$Sigt = P(|t| > t_{calc}).$$

Utilizând tabela Student, de exemplu, pentru o valoare calculată egală cu 2,086, un număr de grade de libertate egal cu 20 și un test unilateral, *Sigt* este:

$$P(t > 2{,}086) = 0{,}025 .$$

În R, această probabilitate se obține cu ajutorul comenzii:

```
> pt(2.086, 20, lower=F)
```

R returnează rezultatul:

```
[1] 0,02499818
```

Dacă se utilizează un test bilateral, semnificația testului se obține practic dublând valoarea calculată mai sus.

Decizia de a accepta sau de a respinge ipoteza nulă, pe baza semnificației testului, presupune următoarele două situații:

- dacă semnificația testului este egală sau mai mare decât pragul de semnificație, $Sigt \geq \alpha$, nu se respinge ipoteza nulă;
- dacă $Sigt < \alpha$, se respinge ipoteza nulă, cu probabilitatea $(1-\alpha)$.

3.5.2. Testarea parametrilor modelului

Parametrii modelului de regresie liniară se testează cu ajutorul testului Student, numit și testul *t*.

Vom exemplifica etapele testării pentru parametrul β_1.

Testul Student (t)

Considerăm un test bilateral, care are următoarele etape:

1. Formularea ipotezelor

$H_0: \beta_1 = 0$ (între cele două variabile nu există o legătură liniară sau variabila independentă nu are o influență liniară semnificativă asupra celei dependente);

$H_1: \beta_1 \neq 0$ (între variabile există o legătură de tip liniar semnificativă statistic).

Observație

În cazul unui test unilateral, ipoteza alternativă se poate formula astfel:

$H_1: \beta_1 > 0$ sau

$H_1: \beta_1 < 0$.

2. Alegerea pragului de semnificație α

De regulă, se ia valoarea *0,05* (în principalele programe de statistică aceasta este valoarea implicită, dar poate fi modificată de utilizator).

3. Alegerea statisticii test

Se alege statistica Student, care are relația:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}}.$$

Observație

Statistica Student depinde de parametrul supus testării, de aceea procedeul de testare pornește de la presupunerea că ipoteza nulă este adevărată.

4. *Determinarea valorii teoretice a testului*

Dacă se acceptă ipoteza nulă, statistica test devine:

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t(n-2)\text{, unde}$$

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{(n-2)\sum_i (x_i - \bar{x})^2}}.$$

Pentru pragul de semnificaţie stabilit, cunoscând legea de repartiţie a statisticii test, pentru $n-2$ grade de libertate, se citeşte din tabela Student valoarea teoretică: $t_{\frac{\alpha}{2};n-2}$.

Se alege valoarea $\alpha/2$, deoarece testul este bilateral (figura 3.6), iar zonele de respingere sunt delimitate de valorile:

$-t_{\frac{\alpha}{2};n-2}$ şi $+t_{\frac{\alpha}{2};n-2}$.

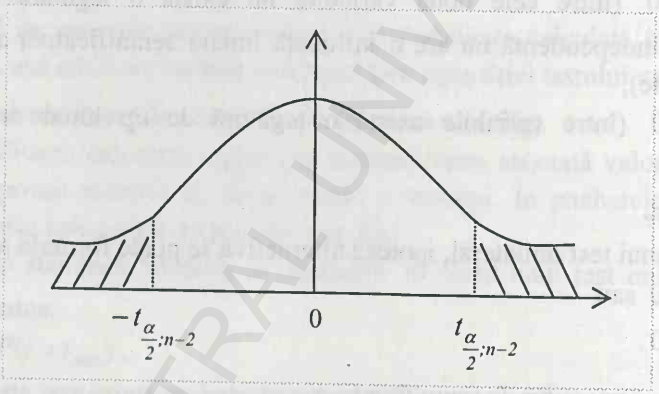


Figura 3.6. *Valorile teoretice ale statisticii Student pentru un nivel de încredere de* $(1-\alpha)$

De exemplu, pentru un prag de semnificaţie de *0,05* şi un eşantion de volum *n=150*, din tabelă se citeşte valoarea $t_{0,025;148} = 1{,}96$. Pe axa reală, valorile -1,96, respectiv 1,96 delimitează regiunile de acceptare şi de respingere a ipotezei nule.

În figura 3.6, ariile haşurate indică probabilităţile ce corespund celor două regiuni de respingere $(-\infty; -t_{\frac{\alpha}{2};n-2})$, respectiv $(t_{\frac{\alpha}{2};n-2}; +\infty)$, iar suma celor două arii este egală cu pragul de semnificaţie α.

Aria nehaşurată este egală cu nivelul de încredere $(1-\alpha)$ şi corespunde regiunii de acceptare, care este dată de intervalul:

BCU IASI / CENTRAL UNIVERSITY LIBRARY

$(-t_{\frac{\alpha}{2};n-2};\ +t_{\frac{\alpha}{2};n-2})$.

5. Determinarea valorii calculate a testului

La nivelul unui eşantion, se obţine o estimaţie a statisticii test:

$$t_{calc}=\frac{b_1}{s_{\hat{\beta}_1}}=\frac{b_1}{\sqrt{\frac{\sum_i (y_i-b_0-b_1x_i)^2}{(n-2)\sum_i (x_i-\bar{x})^2}}}=\frac{b_1}{\sqrt{\frac{\sum_i e_i^2}{(n-2)\sum_i (x_i-\bar{x})^2}}}.$$

Având la bază rezultatele din programele de statistică, relaţia de mai sus se poate scrie astfel:

$$t_{calc}=\frac{Estimate}{Std.\ Error},$$

unde estimaţiile parametrilor modelului de regresie sunt la numărător (*Coefficients Estimates*), iar estimaţiile abaterilor standard ale estimatorilor (*Standard Error*) apar la numitor.

6. Regula de decizie

Regula de decizie, pe baza valorii calculate a testului, este următoarea:

- dacă $t_{calc}\in[-t_{\frac{\alpha}{2};n-2},t_{\frac{\alpha}{2};n-2}]$, nu se respinge ipoteza nulă (se poate considera şi formula „se acceptă ipoteza H_0");
- dacă nu se realizează această condiţie, se respinge ipoteza nulă, cu probabilitatea $(1-\alpha)$.

Dacă se ţine cont de semnificaţia testului, regula de decizie este următoarea:

- dacă $Sigt\geq\alpha$, nu se respinge ipoteza nulă.
- dacă $Sigt<\alpha$, se respinge H_0, cu probabilitatea $(1-\alpha)$.

Exemplu

Pentru modelul de regresie construit pe baza datelor din tabelul 3.2, testarea parametrilor este realizată în R. Rezultatele testării sunt prezentate mai jos, în urma aplicării comenzii:

```
> summary(model1)
Call:
lm(formula = Nat ~ Cas)
Residuals:
```

Min	1Q	Median	3Q	Max
-1,8054	-0,5895	-0,1295	0,6039	1,7151

BCU IASI / CENTRAL UNIVERSITY LIBRARY

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5,7977	0,8430	6,877	2,78e-08 ***
Cas	0,5784	0,1658	3,489	0,0012 **

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Valoarea calculată a testului Student, pentru fiecare parametru, se obține prin relația:

$$t_{calc} = \frac{b_j}{s_{\hat{\beta}_j}}, j = \overline{0,1} \text{ sau } t_{calc} = \frac{Estimate}{Std.\ Error}.$$

Conform rezultatelor prezentate mai sus, valoarea calculată a testului este prezentată în coloana numită *t value* și se obține prin raportul dintre valorile coloanei *Estimate* și *Std. Error*.

De exemplu, pentru parametrul β_0, valoarea statisticii test este:

$$t_{calc} = \frac{5,79}{0,84} = 6,87.$$

Pentru o probabilitate de 0,95 și 40 de grade de libertate, din tabela Student se citește valoarea teoretică $t_{\alpha/2;n-2} = 1,96$. Aplicând regula de decizie, se poate observa că pentru ambii parametri se respinge ipoteza nulă cu nivelul de încredere specificat.

În coloana intitulată $Pr(>|t|)$, sunt prezentate valorile calculate ale probabilităților cu care se obțin cele două estimații ale parametrilor. Este vorba despre semnificația testului pentru fiecare parametru în parte (*Sig t*). Astfel:

- pentru parametrul β_0, $Sigt = 2,78e\text{-}08$;
- pentru β_1, $Sigt = 0,0012$.

Aplicând regula de decizie prin compararea pragului de semnificație cu valoarea *Sig t*, se ia decizia de a respinge ipoteza nulă, cu o probabilitate de 95%, pentru ambii parametri.

3.5.3. Testarea modelului de regresie

Modelul de regresie se testează cu ajutorul testului Fisher. Este un test ce evaluează semnificația modelului de regresie utilizat și apreciază cât de bine modelul aproximează realitatea. Etapele testării sunt prezentate mai jos.

1. *Formularea ipotezelor*

În acest caz, ipoteza nulă se formulează asupra ambilor parametri ai modelului:

$$H_0: \beta_0 = 0, \beta_1 = 0$$

(modelul nu este semnificativ sau variabila independentă nu explică semnificativ variația variabilei independente prin intermediul modelului ales);

$H_1 : \beta_1 \neq 0$

(modelul explică semnificativ legătura dintre variabile).

2. Se fixează pragul de semnificație α

3. Alegerea statisticii test

Pentru această ipoteză, se consideră statistica Fisher. Aceasta se construiește pe baza procedeului de descompunere a variației totale (V_T) a variabilei dependente în două componente: variația explicată (V_E) și variația reziduală (V_R). Utilizând estimatorii componentelor variației, se construiește statistica:

$$F = \frac{\frac{\hat{V}_E}{k-1}}{\frac{\hat{V}_R}{n-k}} = \frac{\hat{V}_E}{\hat{V}_R} \cdot \frac{n-k}{k-1} \sim \Im(k-1, n-k).$$

Această variabilă urmează o lege de repartiție Fisher cu $(k-1)$, respectiv $(n-k)$ grade de libertate, adică este determinată de: k, numărul parametrilor din model (pentru modelul liniar simplu, $k=2$) și n, volumul eșantionului.

4. *Determinarea valorii teoretice a testului*

Pentru un prag de semnificație fixat, se citește valoarea teoretică $F_{\alpha;k-1;n-k}$, ce permite identificarea celor două regiuni: de acceptare și de respingere a ipotezei nule, conform figurii 3.7. Regiunea de respingere corespunde probabilității egale cu pragul de semnificație (regiunea hașurată) și este dată de intervalul:

$(F_{\alpha;k-1;n-k}\ ;+\infty)$.

5. *Valoarea calculată* a statisticii Fisher este:

$$F_{calc} = \frac{\frac{ESS}{k-1}}{\frac{RSS}{n-k}} = \frac{\sum_i (b_0 + b_1 x_i - \bar{y})^2}{\sum_i (y_i - b_0 - b_1 x_i)^2} \cdot \frac{n-k}{k-1}.$$

6. *Decizia* se ia prin compararea valorii calculate a testului cu valoarea din tabela Fisher:

- dacă $F_{calc} > F_{\alpha;k-1;n-k}$, se respinge ipoteza nulă, cu probabilitatea $(1-\alpha)$;
- dacă $F_{calc} \leq F_{\alpha;k-1;n-k}$, se acceptă ipoteza nulă.

BCU IASI / CENTRAL UNIVERSITY LIBRARY

Ca și în cazul testului Student, decizia se poate lua și pe baza unei reguli construite cu ajutorul semnificației testului:

- dacă $Sig F \geq \alpha$, nu se respinge (se acceptă) ipoteza nulă;
- dacă $Sig F < \alpha$, se respinge H_0, cu probabilitatea $(1-\alpha)$.

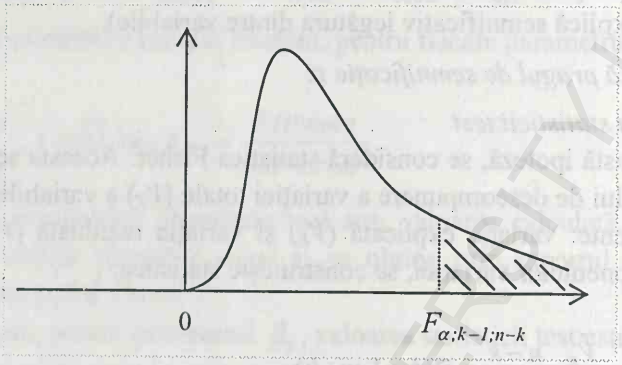


Figura 3.7. *Valoarea teoretică a statisticii Fisher pentru un nivel de încredere de $(1-\alpha)$*

Exemplu

Modelul de regresie estimat pe baza datelor din tabelul 3.2 este testat cu ajutorul testului Fisher, conform rezultatelor obținute în R prin aplicarea comenzii:

```
> anova(model1)
Analysis of Variance Table
Response: Nat
          Df  Sum Sq  Mean Sq  F value    Pr(>F)
Cas        1  10,684  10,6842   12,171  0,001196 ***
Residuals 40  35,115   0,8779
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

În tabelul Anova de mai sus, sunt prezentate estimațiile variației, pe cele două componente (coloana *Sum Sq*), precum și estimațiile varianțelor (coloana *Mean Sq*) obținute prin raportarea acestora la numărul de grade de libertate (coloana *Df*).

Componentele variației sunt:

- variația explicată estimată (apare în dreptul variabilei explicative sau independente), care are valoarea *ESS = 10,684*;
- variația reziduală estimată, care este *RSS = 35,115*.

Gradele de libertate asociate sunt:

- pentru variabila independentă, care determină componenta explicată: $k - 1 = 1$, unde $k = 2$;

- pentru componenta reziduală, $n-k=n-2=40$, unde $n=42$.

Varianţa estimată a erorilor este:

$$s^2=\frac{\sum_i (y_i-b_0-b_1x_i)^2}{n-2}=\frac{35,115}{42-2}=0,877.$$

Valoarea calculată a statisticii Fisher este:

$$F_{calc}=\frac{\frac{ESS}{k-1}}{\frac{RSS}{n-k}}=\frac{10,684}{0,877}=12,17.$$

Valoarea ridicată a statisticii este determinată de valoarea scăzută a estimaţiei varianţei erorilor, ceea ce înseamnă că modelul este valid sau este semnificativ pentru a explica legătura dintre cele două variabile.

Pentru a aplica regula de decizie, se citeşte din tabela Fisher valoarea $F_{\alpha;k-1;n-k}=F_{0,05;1;40}=4,085$. Rezultă că $F_{calc}>F_{\alpha;k-1;n-k}$, adică se respinge ipoteza nulă, cu probabilitatea $(1-\alpha)$.

De asemenea, rezultatele obţinute în R oferă valoarea semnificaţiei testului, $SigF$, care apare în coloana $Pr(>F)$. Deoarece $SigF=0,0011<0,05$, se obţine acelaşi rezultat, adică decizia de a respinge ipoteza nulă.

3.5.4. Testarea indicatorilor de corelaţie

Pentru indicatorii de corelaţie specifici modelării econometrice, se aplică demersul testării statistice prezentat mai sus.

Testarea coeficientului de corelaţie

Coeficientul de corelaţie se testează cu ajutorul unui test Student, prin parcurgerea următoarelor etape:

1. Ipoteze

$H_0:\rho=0$ (între variabile nu există o legătură liniară semnificativă);

$H_1:\rho\neq 0$ (variabilele sunt corelate liniar semnificativ).

2. Pragul de semnificaţie ($\alpha=0,05$)

3. Testul statistic

Se utilizează statistica Student, care în condiţiile acceptării ipotezei nule este:

$$t=\frac{\hat{\rho}}{\sqrt{\frac{1-\hat{\rho}^2}{n-2}}}\sim t(n-2).$$

BCU IASI CENTRU UNIVERSITY LIBRARY

4. Valorile teoretice din tabela Student

Pentru un test bilateral, se citește valoarea $t_{\alpha/2;n-2}$.

5. Valoarea calculată a testului

La nivelul unui eșantion, se calculează:

$$t_{calc} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}.$$

6. Regula de decizie

- dacă $t_{calc} \in [-t_{\alpha/2;n-2}, t_{\alpha/2;n-2}]$, nu se respinge H_0;
- dacă nu se realizează această condiție, se respinge ipoteza nulă, cu probabilitatea $(1-\alpha)$.

Pe baza semnificației testului, se formulează regulile de decizie:

- dacă $Sig\,t \geq \alpha$, nu se respinge ipoteza nulă;
- dacă $Sig\,t < \alpha$, se respinge H_0, cu probabilitatea $(1-\alpha)$.

Exemplu

Pentru modelul de regresie construit pe baza datelor din tabelul 3.2, s-a estimat și testat coeficientul de corelație Pearson utilizând R. Comenzile sunt următoarele:

```
> cor(Nat, Cas)
[1] 0,482
> cor.test(Nat,Cas)
Pearson's product-moment correlation
data: Nat and Cas
t = 3,4886,   df = 40,   p-value = 0,001196
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
0,2098713;  0,6861942
sample estimates:
cor:
0,4829942
```

Deoarece $r = 0{,}483$, iar $n = 42$, rezultă că:

$$t_{calc} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0{,}483}{\sqrt{\frac{1-0{,}233}{40}}} = 3{,}48.$$

BCU "ASE" CENTRAL UNIVERSITY LIBRARY

Din tabela Student, pentru un nivel de încredere de 0,95 şi 40 de grade de libertate, se citeşte: $t_{\alpha/2;n-2} = 1{,}96$.

Comparând cele două valori, se ajunge la decizia de a respinge ipoteza H_0, cu o probabilitate de 0,95, şi se poate considera că legătura liniară directă dintre cele două variabile are o intensitate semnificativă statistic.

Testarea raportului de corelaţie

Demersul testării este prezentat prin etapele de mai jos.

1. Se formulează ipotezele:

$H_0 : \eta = 0$ (între variabile nu există o legătură semnificativă);

$H_1 : \eta > 0$ (variabilele sunt corelate semnificativ).

2. Se alege pragul de semnificaţie α

3. Se utilizează o *statistică Fisher*, care are următoarea expresie:

$$F = \frac{\hat{\eta}^2}{1-\hat{\eta}^2} \cdot \frac{n-k}{k-1},$$

şi urmează o lege de repartiţie Fisher de $k-1$ şi $n-k$ grade de libertate.

4. Se citeşte *valoarea teoretică* $F_{\alpha;k-1;n-k}$ din tabela Fisher, pentru un prag de semnificaţie α stabilit şi pentru $k-1$, respectiv $n-k$ grade de libertate.

5. Se obţine *valoarea calculată* a testului:

$$F_{calc} = \frac{R^2}{1-R^2} \cdot \frac{n-k}{k-1},$$

unde R^2 este raportul de determinaţie calculat la nivelul unui eşantion.

6. Se ia *decizia* pe baza următoarei reguli: dacă $F_{calc} > F_{\alpha;k-1;n-k}$, se respinge ipoteza H_0. În funcţie de semnificaţia testului, dacă $Sig F < \alpha$, se respinge H_0, cu o probabilitate egală cu $(1-\alpha)$.

Observaţie

Testul Fisher utilizat pentru testarea modelului este identic cu cel folosit la testarea raportului de corelaţie:

$$F_{calc} = \frac{ESS}{RSS} \cdot \frac{n-k}{k-1} = \frac{R^2}{1-R^2} \cdot \frac{n-k}{k-1}.$$

La baza acestei egalităţi, stau relaţiile:

$$R^2 = \frac{ESS}{TSS},$$

$TSS = ESS + RSS$.

Exemplu

În R, prin utilizarea comenzii > summary(model1), se obțin atât estimația coeficientului de determinație, cât și testul Fisher aferent:

Multiple R-squared: 0,2333, Adjusted R-squared: 0,2141
F-statistic: 12,17 on 1 and 40 DF, p-value: 0,001196

$$F_{calc} = \frac{R^2}{1-R^2} \cdot \frac{n-k}{k-1} = \frac{0,2333}{1-0,2333} \frac{40}{1} = 12,17 .$$

Pe baza semnificației testului (*p-value = 0,0011*), care este mai mică decât 0,05, se poate decide respingerea ipotezei nule.

3.6. Probleme specifice modelului de regresie liniară simplă

Modelarea unui fenomen cu ajutorul modelului de regresie liniară simplă aduce în discuție o serie de probleme, dintre care ne oprim asupra regresiei prin origine și a liniarității legăturii dintre două variabile.

3.6.1. Regresia prin origine

În procesul de modelare cu ajutorul modelului de regresie liniară simplă, $Y = \beta_0 + \beta_1 X + \varepsilon$, există posibilitatea ca în urma testării parametrilor modelului să se obțină un rezultat care nu permite respingerea ipotezei nule pentru parametrul β_0, iar pentru panta dreptei de regresie, să se constate că parametrul este semnificativ statistic pentru o probabilitate fixată. Într-o asemenea situație, suntem tentați să reconstruim modelul de regresie fără ordonata la origine, adică să estimăm un model de forma $Y = \beta_1 X + \varepsilon$, a cărui dreaptă de regresie trece prin origine.

De asemenea, există situații în care considerentele teoretice impun estimarea unui model care trece prin origine. Această condiție presupune ca lipsa influenței variabilei independente $(X = 0)$ să conducă la o medie zero pentru variabila dependentă.

În cazul modelului de regresie $Y = \beta_1 X + \varepsilon$, aplicarea metodei celor mai mici pătrate se simplifică. Problema de minim care trebuie rezolvată este de forma:

$$S = \sum_i (y_i - \hat{\beta}_1 x_i)^2 = \min .$$

Condiția de extrem conduce la ecuația:

$$2\sum_i (y_i - \hat{\beta}_1 x_i)(-x_i) = 0 \text{ sau}$$

$$\hat{\beta}_1 \sum_i x_i^2 = \sum_i y_i x_i \text{ sau } \hat{\beta}_1 = \frac{\sum_i y_i x_i}{\sum_i x_i^2}.$$

Se poate demonstra că acest estimator este nedeplasat, ca şi în cazul modelului care admite ordonata la origine. Însă utilizarea practică a acestui model întâmpină cel puţin două probleme[3] majore.

Prima este legată de nerespectarea condiţiei $\sum_i \hat{\varepsilon}_i = 0$. A doua problemă este aceea că, pentru modelul de regresie prin origine, R^2 poate fi negativ.

Aşa cum a fost definit pentru modelul de regresie cu ordonata la origine, coeficientului de determinaţie are relaţia:

$$R^2 = 1 - \frac{RSS}{TSS}.$$

În condiţiile în care $\beta_0 = 0$, $RSS = \sum_i e_i^2 = \sum_i y_i^2 - b_1^2 \sum_i x_i^2$.

Deoarece $TSS = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n \cdot \bar{y}^2$, atunci pentru modelul prin origine există şanse ca $RSS > TSS$ şi $R^2 < 0$. Acest lucru se poate întâmpla dacă:

$$b_1^2 \sum_i x_i^2 < n \cdot \bar{y}^2.$$

Ca alternativă pentru R^2, literatura de specialitate a propus un alt indicator pentru modelul de regresie prin origine:

$$r^2 = \frac{(\sum_i y_i x_i)^2}{\sum_i x_i^2 \sum_i y_i^2}, \text{ unde } 0 < r^2 < 1.$$

Observaţia 1

Problemele pe care le ridică modelul de regresie prin origine recomandă pentru practica modelării econometrice evitarea eliminării parametrului β_0 din ecuaţia de regresie, chiar dacă testul Student ar indica această soluţie.

Observaţia 2

Există un caz de regresie prin origine în care problemele menţionate mai sus dispar. Este vorba despre regresia liniară simplă în care variabilele sunt standardizate. În acest caz însă, aşa cum am arătat în secţiunea precedentă cu

[3] A se vedea D. Gujarati, *Basic Econometrics*, McGraw-Hill, 2004, p. 167.

BCU CENTRAL UNIVERSITY LIBRARY

privire la coeficienții de corelație, panta dreptei de regresie este identică cu coeficientul de corelație Pearson și nu mai este necesară calcularea unul alt coeficient.

3.6.2. Liniaritatea și limitele modelului liniar simplu

Să considerăm din nou exemplul prezentat la începutul capitolului 1, al relației dintre nivelul de educație și salariul unui angajat la nivelul unei firme.

Liniaritatea relației dintre cele două variabile presupune o creștere constantă a salariului pentru orice creștere cu o unitate a educației, indiferent dacă această creștere se realizează la nivelul valorilor mici sau la nivelul valorilor mari ale variabilei. Această viteză constantă de creștere este tocmai parametrul $\beta_1 = dY\big|_{dX=1}$ din ecuația liniară de regresie: $Y = \beta_0 + \beta_1 X + \varepsilon$.

În practică, o asemenea situație s-ar traduce, de exemplu, printr-o majorare de salariu cu aceeași valoare absolută β_1 atât pentru un angajat care realizează o creștere a nivelului de educație de la 8 la 9 ani, cât și pentru un angajat care realizează o creștere a educației de la 16 la 17 ani.

Această situație de liniaritate nu corespunde însă cu realitatea. Așa cum se observă și din figura 3.8, pentru intervalul 8 - 12 ani, viteza de creștere a salariului are o anumită valoare, iar pentru intervalul 12 - 21 ani, are o valoare mai mare, așa cum este și normal. Această schimbare a vitezei de variație a variabilei dependente, după un anumit prag valoric pentru variabila independentă, sugerează existența unei legături neliniare între variabile.

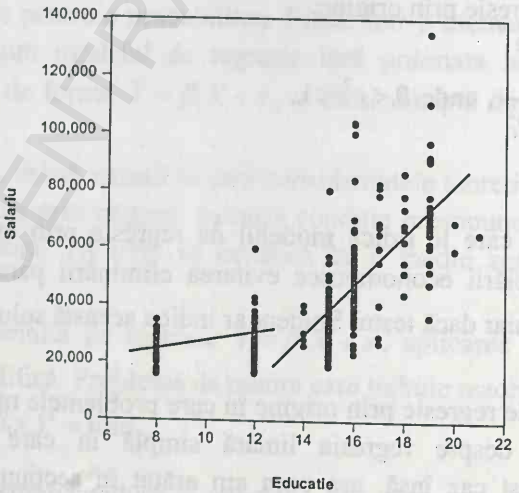


Figura 3.8. *Repartiția angajaților după salariu și nivel de educație*
Sursa: Realizată de autor (date convenționale)

Exemplul prezentat ne atrage atenţia asupra utilizării practice a modelului de regresie liniară simplă. Există situaţii, aşa cum este cazul funcţiei de consum, în care acest model are o acoperire teoretică, iar aplicarea unui model de acest tip nu reprezintă o problemă. Alteori, modelarea cu ajutorul modelului liniar se poate realiza pe baza sugestiilor oferite de analiza grafică şi logică a legăturii dintre variabilele considerate în analiză. Însă, în practica economică şi socială, există relativ puţine cazuri în care între două variabile să identificăm o legătură liniară. De asemenea, există puţine cazuri în care un fenomen, reprezentat de o variabilă dependentă, să fie explicat doar de un singur factor sau o singură variabilă independentă.

Pe lângă dimensiunea practică, importanţa modelului liniar simplu, cu întreaga problematică prezentată în acest capitol, se evidenţiază îndeosebi în plan metodologic. Aspectele analizate în acest capitol reprezintă baza metodologică pentru capitolele următoare în care vor fi analizate alte două categorii de modele: cele liniare multiple, care reprezintă o generalizare a modelului liniar simplu şi o clasă de modele neliniare care pot fi estimate cu ajutorul metodei celor mai mici pătrate, ca urmare a unui procedeu de liniarizare a acestora cu ajutorul unor operatori matematici.

3.7. Studiu de caz

Pentru exemplificarea problemelor metodologice ale modelării econometrice în cazul modelului de regresie liniară simplă, considerăm datele privind populaţia şcolară din învăţământul superior şi Produsul Intern Brut, pe judeţe ale României, în anul 2015 (tabelul 3.3).

Pentru construirea unui model de regresie liniară între cele două variabile, se vor parcurge următoarele etape:

- analiza problemei, care are ca obiectiv delimitarea tipului fiecărei variabile şi a tipului şi sensului legăturii dintre variabile;
- analiza grafică a datelor disponibile, pentru a verifica ipotezele emise în etapa anterioară;
- estimarea parametrilor modelului de regresie;
- testarea parametrilor şi a modelului obţinut;
- utilizarea în practică a modelului pentru predicţii.

a. Analiza problemei

Analiza logică sugerează că populaţia şcolară din învăţământul liceal, la nivelul judeţelor României (exceptând Bucureşti), în anul 2015, este variabila explicată sau variabila dependentă.

Tabelul 3.3. *Repartiția județelor României după populația școlară din învățământul liceal și PIB, în anul 2015*

Județul	Populația școlară în înv. sup. (*Y*)	PIB (*X*, mil. lei)
Bacău	19899	12836,5
Botoșani	14518	6629,9
Iași	25522	20604,2
Neamț	17694	8843,0
Suceava	26538	12249,3
Vaslui	15185	6006,5
Brăila	10138	7357,1
Buzău	14431	9431,1
Constanța	25215	33901,5
Galați	18951	12649,1
Tulcea	6375	5151,6
Vrancea	10829	6748,9
Argeș	22096	17878,3
Călărași	8182	5845,4
Dâmbovița	15449	11594,5
Giurgiu	6397	6027,8
Ialomița	8175	6161,4
Prahova	24477	32602,7
Teleorman	12751	6704,6
Dolj	21589	16563,3
Gorj	15294	9980,6
Mehedinți	9952	4748,3
Olt	16282	8459,5
Vâlcea	14068	8588,3
Arad	14106	13820,0
Caraș-Severin	10754	6831,8
Hunedoara	15329	10917,1
Timiș	22139	29770,5
Bihor	20322	15289,6
Bistrița-Năsăud	10381	6779,1
Cluj	20259	29805,6
Maramureș	17206	10893,6
Satu Mare	10129	8106,8
Sălaj	7708	5766,8
Alba	12741	11009,3
Brașov	15044	22014,4
Covasna	5613	4723,7
Harghita	9942	7104,3
Mureș	16245	14445,9
Sibiu	11746	13907,4
Ilfov	6815	17149,5

Sursa: Realizat după Anuarul Statistic al României, 2016, www.insse.ro

BCU IASI CENTRALA UNIVERSITARA LIBRARY

Variabila explicată este într-o relaţie de dependenţă directă cu Produsul Intern Brut (variabila independentă), care exprimă sintetic gradul de dezvoltare economică al fiecărui judeţ.

Ideea unei legături liniare directe între cele două variabile presupune că o creştere a PIB-ului de la un judeţ la altul va determina o creştere a populaţiei şcolare din învăţământul superior.

b. Analiza grafică

Analiza statistică a datelor înregistrate în anul 2014 cu privire la cele două variabile se realizează cu ajutorul unei reprezentări grafice bidimensionale de tip „nor de puncte". Diagrama din figura 3.9 sugerează existenţa unei legături liniare directe între populaţia şcolară din învăţământul superior şi PIB, la nivelul judeţelor României.

Realizăm reprezentarea grafică în R pe baza comenzilor:

```
> plot(Y~X)
> abline(lm(Y~X))
```

Punctele empirice de forma $A_i(x_i, y_i)$ se repartizează în jurul unei drepte:

$Y = \beta_0 + \beta_1 X$.

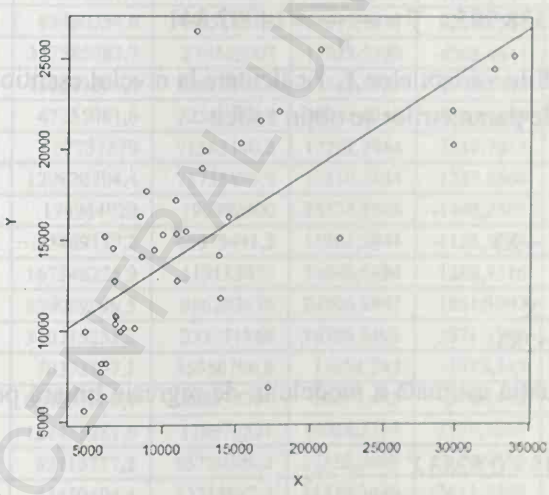


Figura 3.9. *Aprecierea statistică a dependenţei dintre cele două variabile*
Sursa: Realizată de autor potrivit datelor din tabelul 3.3

Pentru variabilele considerate, se poate propune un model econometric de regresie liniară simplă:

$Y = \beta_0 + \beta_1 X + \varepsilon$.

BCU IASI / CENTRAL UNIVERSITY LIBRARY

c. Estimarea parametrilor modelului

Estimarea parametrilor modelului se poate realiza prin calcul clasic, pe baza relațiilor de calcul ale estimațiilor obținute prin metoda celor mai mici pătrate. De asemenea, calculul se poate realiza și cu ajutorul unui software specializat de statistică, așa cum este R, sau cu ajutorul programului Excel, care dispune de un modul de analiză statistică a datelor.

Calculul clasic se realizează pe baza datelor din tabelul 3.3. Elementele de calcul sunt prezentate în tabelul 3.4.

Estimarea punctuală

Estimațiile parametrilor se calculează pe baza relațiilor:

$$b_1 = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} = \frac{41 \cdot 8749187956 - 505899 \cdot 606486}{41 \cdot 8638008652 - 505899^2} = 0{,}52834,$$

$$b_0 = \bar{y} - b_1\bar{x} = 14792{,}341 - 0{,}52834 \cdot 12338{,}995 = 8273{,}144,$$

unde:

$$\bar{x} = \frac{\sum_i x_i}{n} = 12338{,}995; \quad \bar{y} = \frac{\sum_i y_i}{n} = 14792{,}341$$

reprezintă mediile variabilelor X, Y calculate la nivelul eșantionului.

În R, estimațiile parametrilor se obțin astfel:

```
> lm(Y~X)

Call:
lm(formula = Y ~ X)
Coefficients:
(Intercept)        X
8273,1445    0,5283
```

Rezultă că ecuația estimată a modelului de regresie liniară pentru cele două variabile este:

$$Y_X = 8273{,}1445 + 0{,}5283X.$$

Estimarea prin interval de încredere

Se realizează cu ajutorul relațiilor:

$b_1 \pm t_{\alpha/2} s_{\hat{\beta}_1}$, respectiv

$b_0 \pm t_{\alpha/2} s_{\hat{\beta}_0}$.

Estimațiile pentru abaterile standard ale estimatorilor sunt calculate pe baza datelor din tabelul 3.3.

BCU "LUCIAN BLAGA" CENTRAL UNIVERSITY LIBRARY

Tabelul 3.4. *Elemente de calcul*

	y_i	x_i	$x_i y_i$	x_i^2	y_{x_i}	e_i	e_i^2	$(x_i - \bar{x})^2$
1	19899	12836,5	255433513,5	164775732	15054,6675	4844,3326	23467557,85	247511,225
2	14518	6629,9	96252888,2	43955574	11775,7207	2742,2793	7520095,924	32593765,72
3	25522	20604,2	525860392,4	424533058	19158,3434	6363,6566	40496125,83	68313613,69
4	17694	8843	156468042	78198649	12944,9014	4749,0986	22553937,51	12221981,04
5	26538	12249,3	325071923,4	150045350	14744,4497	11793,55	139087828,9	8045,193025
6	15185	6006,5	91208702,5	36078042,3	11446,3785	3738,6216	13977291,09	40100492,93
7	10138	7357,1	74586279,8	54126920,4	12159,9004	-2021,9004	4088081,349	24819277,79
8	14431	9431,1	136100204,1	88945647,2	13255,5946	1175,4054	1381577,784	8455853,331
9	25215	33901,5	854826322,5	1149311702	26183,307	-968,30695	937618,3494	464941621,9
10	18951	12649,1	239713094,1	159999731	14955,664	3995,336	15962709,51	96165,11102
11	6375	5151,6	32841450	26538982,6	10994,7348	-4619,7348	21341949,44	51658646,89
12	10829	6748,9	73083838,1	45547651,2	11838,5884	-1009,5884	1019268,677	31249162,11
13	22096	17878,3	395038916,8	319633611	17718,2504	4377,7496	19164691,65	30683899,88
14	8182	5845,4	47827062,8	34168701,2	11361,2693	-3179,2693	10107753,41	42166776,02
15	15449	11594,5	179123430,5	134432430	14398,5189	1050,4812	1103510,647	554272,805
16	6397	6027,8	38559836,6	36334372,8	11457,6312	-5060,6312	25609988,55	39831182,33
17	8175	6161,4	50369445	37962850	11528,2121	-3353,2121	11244031,52	38162679,98
18	24477	32602,7	798016287,9	1062936047	25497,1509	-1020,1509	1040707,879	410617740,3
19	12751	6704,6	85490354,6	44951661,2	11815,1847	935,81532	875750,3131	31746407,02
20	21589	16563,3	357585083,7	274342907	17023,5359	4565,4641	20843462,54	17844752,73
21	15294	9980,6	152643296,4	99612376,4	13545,8955	1748,1045	3055869,413	5562026,976
22	9952	4748,3	47255081,6	22546352,9	10781,6714	-829,67139	688354,6154	57618650,58
23	16282	8459,5	137737579	71563140,3	12742,2984	3539,7017	12529487,77	15050481,46
24	14068	8588,3	120820204,4	73758896,9	12810,3434	1257,6566	1581700,149	14067712,98
25	14106	13820	194944920	190992400	15574,2505	-1468,2505	2155759,531	2193375,81
26	10754	6831,8	73469177,2	46673491,2	11882,3844	-1128,3844	1273251,444	30329196,77
27	15329	10917,1	167348225,9	119183072	14040,6484	1288,3516	1659849,768	2021785,391
28	22139	29770,5	659089099,5	886282670	24000,8997	-1861,8997	3466670,307	303857366,6
29	20322	15289,6	310715251,2	233771868	16350,6402	3971,3598	15771698,82	8706069,866
30	10381	6779,1	70373837,1	45956196,8	11854,543	-1473,543	2171329,061	30912432,41
31	20259	29805,6	603831650,4	888373791	24019,443	-3760,443	14140931,41	305082290,2
32	17206	10893,6	187435281,6	118670521	14028,2334	3177,7666	10098200,69	2089166,706
33	10129	8106,8	82113777,2	65720206,2	12555,9669	-2426,9669	5890168,528	17911474,52
34	7708	5766,8	44450494,4	33255982,2	11319,7449	-3611,7449	13044701,51	43193747,12
35	12741	11009,3	140269491,3	121204686	14089,3577	-1348,3577	1818068,46	1768088,793
36	15044	22014,4	331184633,6	484633807	19903,352	-4859,352	23613302,05	93613461,91
37	5613	4723,7	26514128,1	22313341,7	10768,6752	-5155,6752	26580986,87	57992717,94
38	9942	7104,3	70630950,6	50471078,5	12026,3462	-2084,3462	4344499,04	27402031,74
39	16245	14445,9	234673645,5	208684027	15904,9135	340,08653	115658,8479	4439048,679
40	11746	13907,4	163356320,4	193415775	15620,4239	-3874,4239	15011160,71	2459894,244
41	6815	17149,5	116873842,5	294105350	17333,2254	-10518,225	110633064,5	23140958,36
Total	**606486**	**505898,8**	**8749187956**	**8638008652**	-	-	**651468652,26**	**2395725827**

Sursa: Realizat de autor potrivit datelor din tabelul 3.3

$$s = \sqrt{\frac{\sum_i e_i^2}{(n-2)}} = \sqrt{\frac{651468652,2}{39}} = 4087.,$$

$$s_{\hat{\beta}_1} = \sqrt{\frac{s^2}{\sum_i (x_i - \bar{x})^2}} = \sqrt{\frac{4087^2}{2395725827}} = 0,0835,$$

$$s_{\hat{\beta}_0} = \sqrt{s^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2})} = \sqrt{4087^2(\frac{1}{41} + \frac{152250800,6}{2395725827})} = 1212,022.$$

Pentru un nivel de încredere de 0,95, se citeşte din tabela Student valoarea teoretică $t_{0,025;39} = 2,022$. În R, valoarea se obține cu ajutorul comenzii:

```
> qt(0.975, 39)
[1] 2,022691
```

Pe baza datelor de mai sus, se calculează intervalele de încredere pentru cei doi parametri:

- pentru parametrul $\beta_1 : (b_1 \pm t_{\alpha/2} s_{\hat{\beta}_1})$, adică

  $(0,5283 \pm 2,0226 \cdot 0,0835)$ sau *(0,359; 0,697)*;
- pentru parametrul $\beta_0 : (b_0 \pm t_{\alpha/2} s_{\hat{\beta}_0})$, adică

  $(8273,144 \pm 2,0226 \cdot 1212,022)$ sau *(5821,6; 10724,6)*.

În R, intervalele de încredere se obțin astfel:

```
> confint(lm(Y~X))
                2,5%          97,5%
(Intercept) 5821,597051  1,072469e+04
X              0,3594426  6,972393e -01
```

Interpretarea rezultatelor

- deoarece $b_1 > 0$, se confirmă că legătura dintre cele două variabile este una directă;
- estimația $b_1 = 0,5283$ indică o creştere medie a efectivului populației şcolare liceale la nivelul unui județ cu 0,5283 persoane (aproximativ 1 persoană) la o creştere a PIB-ului cu 1 mil. de lei;
- estimația $b_0 = 8273,144$ este o valoare ce nu are un conținut economic practic. Teoretic, valoarea reprezintă efectivul mediu estimat al populației şcolare la nivel de județ pentru un PIB egal cu zero;

- cu o probabilitate de 0,95, parametrul β_0 este acoperit de intervalul *(0,36 ; 0,69)*;
- cu o probabilitate de 0,95, parametrul β_1 este acoperit de intervalul *(5821,6 ; 10724,6)*.

d. Testarea parametrilor modelului

Se parcurg etapele testării pentru testul Student:

1. *Ipoteze*

$H_0 : \beta_1 = 0$;

$H_1 : \beta_1 \neq 0$.

2. Pragul de semnificaţie α ia valoarea *0,05*.

3. Statistica *Student*, în condiţiile acceptării ipotezei nule, este:

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t(n-2).$$

4. *Valorile teoretice* $t_{\alpha/2;n-2}$ se citesc din tabela Student pentru $n-2$ grade de libertate şi $\alpha/2$ (deoarece testul este bilateral). Pentru un prag de semnificaţie de 0,05 şi un eşantion de volum $n=41$, din tabel, se citeşte valoarea $t_{0,025;39} = 2,022$.

5. *Valoarea calculată* la nivelul unui eşantion este:

$$t_{calc} = \frac{b_1}{s_{\hat{\beta}_1}} = \frac{0,5283}{0,0835} = 6,32.$$

6. Decizia

Deoarece $t_{calc} = 6,32 > t_{\alpha/2;n-2} = 2,022$, se respinge ipoteza H_0, cu o probabilitate de 0,95, adică se respinge ipoteza că între cele două variabile nu ar exista o legătură de tip liniar (sau că panta dreptei de regresie nu ar fi semnificativ diferită de zero).

Analog, pentru parametrul β_0, valoarea calculată a testului Student este:

$$t_{calc} = \frac{b_0}{s_{\hat{\beta}_0}} = \frac{8273,144}{1212,022} = 6,82.$$

Deoarece $t_{calc} = 6,82 > t_{\alpha/2;n-2} = 2,022$, se respinge ipoteza nulă şi pentru acest parametru, cu probabilitatea 0,95.

Rezultatele testării în R sunt prezentate în tabelul de mai jos.

BCU IAŞI - CENTRAL UNIVERSITY LIBRARY

```
> summary(lm(Y~X))

Call:
lm(formula = Y ~ X)
Coefficients:
             Estimate     Std. Error    t value     Pr(>|t|)
(Intercept) 8273,1445    1212,0228      6,826    3,72e-08 ***
X             0,5283       0,0835      6,327    1,82e-07 ***
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

Testul Student pentru cei doi parametri indică o valoare *Sigt* foarte mică, apropiată de zero. Deoarece $Sigt < 0,05$, se ia decizia de a respinge ipoteza nulă pentru ambii parametri, cu o probabilitate de 95%.

e. Testarea modelului

Se parcurg etapele clasice ale testării:

- *Formularea ipotezelor*

  $H_0: \beta_0 = 0, \beta_1 = 0;$

  $H_1: \beta_1 \neq 0.$

- *Statistica Fisher*

$$F = \frac{\hat{V}_E}{\hat{V}_R} \cdot \frac{n-k}{k-1} \sim \Im(k-1, n-k).$$

- *Valoarea teoretică a statisticii* este:

  $F_{0,05;1;39} = 4,085.$

- *Valoarea calculată a statisticii Fisher* este:

$$F_{calc} = \frac{\sum_i (b_0 + b_1 x_i - \bar{y})^2}{\sum_i (y_i - b_0 - b_1 x_i)^2} \frac{n-k}{k-1} = \frac{ESS}{RSS} \frac{n-k}{k-1}.$$

Tabelul Anova, obținut în R, ne oferă estimațiile pentru variația explicată și cea reziduală, conform rezultatelor obținute mai jos:

```
> anova(lm(Y~X))

Analysis of Variance Table
Response: Nat
           Df    Sum Sq     Mean Sq    F value     Pr(>F)
Cas         1 668752979   668752979     40,035 1,819e-07 ***
Residuals  39 651468638    16704324
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

BCU "CAROL I" CENTRAL UNIVERSITY LIBRARY

- *Sum of Squares* pentru variabila X este estimaţia variaţiei explicate şi are valoarea *668752979*;
- *Sum of Squares Residuals* este estimaţia variaţiei reziduale şi are valoarea *651468638*.

În coloana *Df*, se prezintă gradele de libertate ale estimatorilor fiecărei componente a variaţiei: $k-1=1$, pentru variaţia explicată, şi $n-k=39$, pentru variaţia reziduală.

În coloana *Mean Sq*, sunt calculate estimaţiile pentru dispersia explicată şi dispersia reziduală, ca raport între estimaţia variaţiei şi numărul corespunzător de grade de libertate.

Valoarea calculată a testului se obţine astfel:

$$F_{calc}=\frac{ESS}{RSS}\cdot\frac{n-k}{k-1}=\frac{668752979}{16704324}=40{,}03.$$

- *Decizia*

În mod clasic, decizia cu privire la ipoteza nulă se ia prin compararea valorii calculate cu valoarea teoretică corespunzătoare care se citeşte din tabela Fisher:

$$F_{\alpha,k-1,n-k}=F_{0{,}05;1;39}=4{,}085.$$

Deoarece $F_{calc}>F_{\alpha,k-1,n-k}$, se respinge ipoteza nulă, cu o încredere de 0,95.

Dacă se consideră semnificaţia testului (*Sig F*), care are valoarea 1,819e-07, se ajunge la aceeaşi concluzie, anume aceea de a respinge ipoteza nulă, cu o probabilitate de 0,95.

f. Testarea raportului de corelaţie

Testul echivalent al testului Fisher, prezentat anterior, este construit pe baza estimatorului coeficientului de determinaţie:

$$F=\frac{\hat{V}_E}{\hat{V}_R}\cdot\frac{n-k}{k-1}=\frac{\hat{\eta}^2}{1-\hat{\eta}^2}\cdot\frac{n-k}{k-1}$$

Acest test este utilizat pentru testarea raportului de corelaţie. Ipotezele testului sunt:

$H_0:\eta=0$;

$H_1:\eta>0$.

În R, estimaţia raportului de corelaţie şi a raportului de determinaţie sunt prezentate mai jos:

Residual standard error: 4087 on 39 degrees of freedom
Multiple R-squared: 0,5065, Adjusted R-squared: 0,4939
F-statistic: 40,03 on 1 and 39 DF, p-value: 1,819e-07

BCU "SPIRU HARET" DIGITAL UNIVERSITY LIBRARY

Valoarea calculată a testului este aceeaşi dacă se aplică relația:

$$F_{calc} = \frac{R^2}{1-R^2} \cdot \frac{n-k}{k-1} = \frac{0,5065}{0,4935} \cdot \frac{39}{1} = 40,03.$$

În concluzie, se ia decizia de a respinge ipoteza nulă şi se consideră că între cele două variabile există o legătură semnificativă.

g. Testarea coeficientului de corelație

Ipotezele pentru acest test sunt:

$H_0: \rho = 0$ (între variabile nu există o legătură liniară semnificativă);

$H_1: \rho \neq 0$ (variabilele sunt corelate semnificativ).

Pentru un prag de semnificație de *0,05*, se citeşte valoarea teoretică din tabela Student: $t_{\alpha/2;n-2} = t_{0,025;39} = 2,022$.

Valoarea calculată a testului se determină pe baza estimației coeficientului de corelație. În R, aceasta se obține:

```
> cor(Y,X)
[1] 0,7117205
```

Rezultă:

$$t_{calc} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,7117}{\sqrt{\frac{1-0,5065}{39}}} = 6,32.$$

Deoarece t_{calc} este mai mare decât *2,022*, se respinge ipoteza nulă, cu o probabilitate de *0,95*.

Utilizând R, coeficientul de corelație se testează astfel:

```
> cor.test(Y,X)

Pearson's product-moment correlation
data: Y and X
t = 6,3273,    df = 39,    p-value = 1,819e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
0,5173494;    0,8362620
sample estimates:
cor:
0,7117205
```

Evident, se obțin aceleași rezultate ca cele prezentate mai sus.

BCU "CENTRAL UNIVERSITY LIBRARY"

3.8. Teste și exerciții

Teste grilă

1. În modelul de regresie liniară simplă, parametrul β_1 reprezintă:

a) nivelul mediu al variabilei dependente dacă variabila independentă ia valoarea 1;

b) ordonata la origine;

c) variația absolută medie a variabilei dependente la o variație absolută cu o unitate a variabilei independente;

d) panta dreptei de regresie.

2. Pentru un model de regresie liniară simplă, coeficientul de corelație este identic cu panta dreptei de regresie dacă:

a) valorile variabilei dependente sunt mai mari decât cele ale variabilei independente;

b) valorile celor două variabile sunt standardizate;

c) valorile celor două variabile sunt diferite.

3. Coeficientul de determinație arată:

a) gradul de intensitate a legăturii dintre două variabile;

b) ponderea variației variabilei dependente explicate de variația variabilei independente;

c) egalitatea mediilor a două populații.

4. Pentru variabilele *rata mortalității* (decese la 1000 de locuitori) și *rata natalității* (născuți vii la 1000 de locuitori), la nivelul județelor României, s-a obținut rezultatul de mai jos.

```
> cor.test(Mort,Nat)

Pearson's product-moment correlation
data:  Mort and Nat
t = -4,4086,   df = 40,    p-value = 7,619e-05
alternative hypothesis:  true mean is not equal to 0
sample estimates:
cor:
-0,5718455
```

Au loc afirmațiile:

a) coeficientul de corelație este semnificativ statistic, cu o probabilitate de 0,95;

b) între variabile există o legătură inversă semnificativă;

c) estimația coeficientului de corelație este egală cu -4,4;

d) numărul de grade de libertate asociat testului Student este egal cu 42-2.

5. Pentru variabilele *nivelul salariului* ($) și *perioada lucrată în companie* (luni), s-a obținut rezultatul de mai jos.

```
> lm(salary~jobtime)

Call:
lm(formula = salary ~ jobtime)
Coefficients:
(Intercept)    jobtime
  22843,3       142,7
```

Este valabilă interpretarea:

a) la o creștere cu 1 an a numărului de luni lucrate, nivelul salariului crește, în medie, cu 22843,3$;

b) la o creștere cu 1 an a numărului de ani de studii, nivelul salariului crește, în medie, cu 142,7$;

c) un angajat care nu a lucrat nici o lună în companie câștigă, în medie, 142,7$;

d) ecuația estimată a modelului de regresie este: $Y_X = 22843,3 + 142,7X$.

6. Pentru variabilele *nivelul salariului* ($) și *perioada lucrată în companie* (luni), s-a obținut rezultatul de mai jos.

```
> summary(lm(salary~jobtime))

Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  22843,32    6362,21     3,590     0,000365***
jobtime        142,72      77,84     1,833     0,067366.
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

Sunt valabile enunțurile:

a) panta dreptei este semnificativ diferită de zero, cu o probabilitate de 0,95;

b) ordonata la origine este semnificativ diferită de zero, cu o probabilitate de 0,999;

c) panta dreptei este semnificativ diferită de zero, cu o probabilitate de 0,9.

7. Pentru variabilele *rata mortalității* (decese la 1000 de locuitori) și *rata natalității* (născuți vii la 1000 de locuitori), la nivelul județelor României, s-a obținut rezultatul de mai jos.

BCU IASI CENTRAL UNIVERSITY LIBRARY

```
> anova(lm(Nat~Mort))
Analysis of Variance Table
Response: Nat
           Df  Sum Sq  Mean Sq  F value     Pr(>F)
Mort        1  14,977  14,9766  19,436  7,619e-05***
Residuals  40  30,822   0,7706
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

Sunt valabile enunțurile:

a) modelul de regresie este semnificativ statistic, cu o probabilitate de 0,999;

b) estimația variației totale este $TSS = 45,799$;

c) estimația coeficientului de determinație este 32,7%;

d) modelul nu este semnificativ statistic.

8. În cazul unui model liniar simplu, au loc:

a) viteza de variația a variabilei dependente în raport cu cea independentă este constantă;

b) $r = R$;

c) coeficientul de corelație este pozitiv;

d) raportul de corelație este pozitiv.

9. Pentru două variabile X, Y, se cunoaște ecuația estimată a modelului de regresie: $Y_X = 2,4 + 0,7X$. Sunt adevărate enunțurile:

a) ecuația este valabilă pentru variabilele *consum* și *venit*;

b) dacă X crește cu 10 unități, Y crește cu 0,7 unități;

c) o creștere medie a lui Y cu 7 unități presupune o creștere a lui X cu 10 unități.

10. Pentru două variabile X, Y, se cunosc: $ESS = 30$, $RSS = 10$, $n = 100$. Atunci:

a) $R^2 = 0,75$;

b) modelul de regresie dintre cele 2 variabile este semnificativ statistic, cu o probabilitate de 0,95;

c) între cele două variabile există o legătură directă.

Exerciții

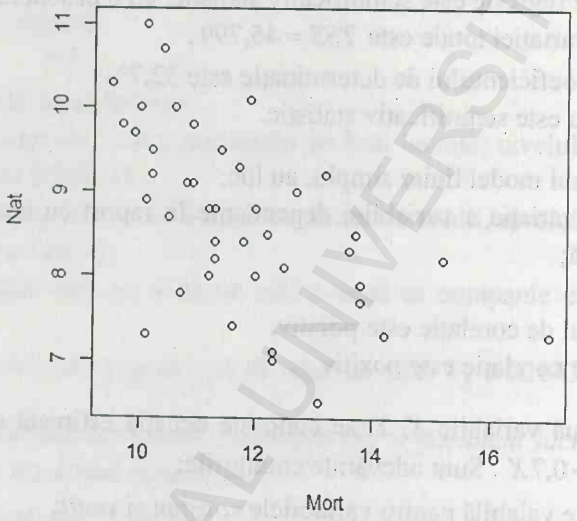
1. Din lista de variabile de mai jos identificați posibile relații de dependență care să se constituie apoi în modele econometrice: vânzările unei companii, consumul unei familii, salariul unui angajat, cheltuielile cu publicitatea ale unei companii, venitul familiei, mărimea familiei, experiența de muncă a unui angajat,

BCU "MIHAIL KOGĂLNICEANU" LIBRARY

nivelul de educație al angajatului, prețul petrolului, vânzările de autoturisme, prețul alimentelor.

2. Care sunt componentele modelului de regresie? Ce rol joacă variabila aleatoare reziduală?

3. Considerând reprezentarea grafică de mai jos pentru 2 variabile X, Y (rata natalității și rata mortalității pe județe), analizați ce tip de model ar explica legătura dintre variabile.



4. Care sunt asemănările și deosebirile dintre modele de mai jos:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 (X - \mu_X) + \varepsilon$$

$$Y = \beta_0 + \mu_X + \beta_1 X + \varepsilon$$

$$Y + \mu_X = \beta_0 + \beta_1 X + \varepsilon$$

5. Pentru 2 variabile, X, Y, se cunosc rezultatele de mai jos privind modelarea econometrică a legăturii dintre acestea.

Response: Y

	Df	Sum Sq
X	1	270
Residuals	100	130

a) calculați și interpretați estimația coeficientului de determinație;

b) testați semnificația coeficientului de determinație cu o probabilitate de

BCU ASUCENTRAL UNIVERSITY LIBRARY

0,95;

c) dacă între variabile există o legătură inversă, estimați valoarea coeficientului de corelație Pearson.

6. Se dau datele din tabelul de mai jos privind ratele de natalitate și divorțialitate, la nivelul județelor României, în anul 2014 (sursa www.insse.ro).

Județul	Nat.	Div.
Bihor	9,3	1,2
Bistrita-Nasaud	10	0,6
Cluj	9,1	0,97
Maramures	8,7	1,32
Satu Mare	9,1	1,26
Salaj	10,1	1,1
Alba	8,2	1,34
Brasov	9,8	1,78
Covasna	9,8	0,71
Harghita	10	0,85
Mures	9,5	1,04
Sibiu	9,7	1,22
Bacau	8,8	1,34
Botosani	9	1,62
Iasi	10,3	1,16
Neamt	8,4	1,37
Suceava	10,7	1,3
Vaslui	8,8	1,4
Braila	6,5	1,95
Buzau	7,7	1,09
Constanta	9,5	1,66

Județul	Nat.	Div.
Galati	7,8	1,18
Tulcea	8,1	0,83
Vrancea	8,8	0,91
Arges	8	1,39
Calarasi	8,5	1,29
Dâmbovita	8,4	1,03
Giurgiu	8,2	0,53
Ialomita	9,2	0,91
Prahova	8	1,33
Teleorman	7,3	1,13
Ilfov	11	1,28
Bucuresti	8,9	1,31
Dolj	8,3	0,63
Gorj	7,4	1,55
Mehedinti	7,9	0,85
Olt	7,3	0,93
Vâlcea	7,3	0,98
Arad	8,5	1,12
Caras-Severin	7	1,66
Hunedoara	7,1	1,55
Timis	9,2	1,1

Se cere:

- să se reprezinte grafic distribuția bidimensională (cu ajutorul graficului de tip scatterplot) folosind R;
- din punct de vedere teoretic există suport pentru a explica legătura dintre cele două variabile cu ajutorul unui model liniar de regresie simplă?
- dacă analizăm reprezentarea grafică, ce putem spune despre o posibilă legătură dintre cele două variabile?

7. În urma modelării econometrice a legăturii dintre variabilele X (experiența anterioară, luni lucrate în altă companie) și Y (salariu), se cunosc rezultatele:

Coefficients:

	Estimate	Std. Error
(Intercept)	35945,029	1060,488
prevexp	-15,913	7,479

BCU IASI / CENTRAL UNIVERSITY LIBRARY

a) scrieți ecuația estimată a modelului de regresie;

b) interpretați estimațiile parametrilor (cum apreciați rezultatul privind panta dreptei de regresie?);

c) calculați valorile testului Student pentru cei doi parametri și luați decizia, cu o probabilitate de 0,95, privind semnificația parametrilor modelului de regresie;

d) calculați intervalele de încredere pentru cei doi parametri. Se poate lua decizia privind semnificația parametrilor dacă se cunosc limitele intervalelor de încredere?

8. Pentru datele de la exercițiul 6, să se construiască în R un model de regresie. Să se verifice rezultatele privind semnificația dreptei de regresie și a coeficientului de determinație. Coroborați rezultatele cu analiza grafică realizată la exercițiul anterior.

BCU IASI CENTRAL UNIVERSITY LIBRARY

Capitolul 4

REGRESIA LINIARĂ MULTIPLĂ

Modelul de regresie liniară multiplă reprezintă o generalizare a modelului de regresie liniară simplă. În model, apar două sau mai multe variabile independente, care sunt considerate factori de influenţă. Variaţia variabilei dependente este explicată prin variaţia simultană a factorilor, după modelul unei ecuaţii liniare.

În acest capitol, vor fi prezentate principalele aspecte metodologice ale modelării econometrice pentru modelul liniar multiplu: prezentarea modelului econometric, estimarea parametrilor, coeficienţii de corelaţie, testarea parametrilor şi a modelului de regresie.

4.1. Prezentarea modelului

Considerăm un model de regresie liniară multiplă care conţine p variabile independente: $X_1, X_2, ..., X_p$. Formal, modelul este dat prin relaţia:

$$Y = E(Y/X) + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon .$$

În model, apar k parametri ($k = p + 1$) şi au următoarea semnificaţie:

- β_0 este valoarea medie a variabilei dependente, în condiţiile în care influenţa variabilelor independente este nulă. În programele de statistică, acest parametru se asociază unei variabile speciale care se numeşte *constanta modelului*, pentru că este coeficientul unei variabile degenerate, $X_0 = 1$.
- $\beta_j = \frac{\partial Y}{\partial X_j}$, $j = \overline{1, p}$, reprezintă variaţia absolută a variabilei dependente la o variaţie absolută cu o unitate a variabilei independente X_j, în condiţiile în care influenţa celorlalte variabile independente este menţinută constantă sau

este controlată influența acestora. Acești parametri se numesc *coeficienți de regresie parțială* și arată influența parțială a fiecărei variabile independente asupra variabilei dependente.

Observație

Pentru a înțelege mai bine semnificația unui parametru într-un model de regresie multiplă, să considerăm parametrul β_1. Acesta măsoară variația variabilei dependente dacă se realizează o modificare ∂X_1 pentru prima variabilă, iar celelalte își păstrează valorile sau nu variază. Altfel spus, β_1 este diferența dintre valoarea așteptată pentru variabila dependentă când variabilele independente iau valorile $x_1 + \partial X_1, x_2, \ldots, x_p$ și valoarea așteptată pentru variabila dependentă când variabilele independente iau valorile $x_1, x_2, \ldots, x_p$.

În general, dacă $\partial X_j = 1$, rezultă $\beta_j = \partial Y, j = \overline{1,p}$.

Prezentarea matriceală

Modelul multiplu admite și o abordare cu ajutorul matricelor. Acesta se poate scrie sub formă matriceală astfel:

$Y = X \cdot \beta + \varepsilon$, unde

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix}, \; X = \begin{pmatrix} X_0 & X_1 & \ldots & X_p \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & \ldots & x_{p1} \\ 1 & x_{12} & x_{22} & \ldots & x_{p2} \\ \ldots & \ldots & \ldots & \ldots & \\ 1 & x_{1n} & x_{2n} & \ldots & x_{pn} \end{pmatrix}, \; \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \ldots \\ \beta_p \end{pmatrix}, \; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{pmatrix},$$

iar

- p este numărul de variabile independente;
- k este numărul de parametri din model;
- n este volumul de date disponibile.

Prima coloană din matricea corespunzătoare valorilor variabilelor independente este coloana variabilei constante (X_0), ale cărei valori sunt egale cu unu.

Pentru $p = 2$ sau $k = 3$, avem modelul de regresie multiplă cel mai simplu, adică modelul cu două variabile independente:

$y_i = E(Y / X_j) + \varepsilon_i = \beta_0 X_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, unde $X_0 = 1$ sau

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$.

Scris matriceal, modelul de mai sus este dat prin:

$Y = X \cdot \beta + \varepsilon$, unde

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}, \quad X = (X_0\ X_1\ X_2) = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ & \dots & \dots \\ 1 & x_{1n} & x_{2n} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Ipotezele modelului clasic de regresie

Considerăm modelul de regresie liniară multiplă cu două variabile independente:

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$.

Ipotezele care trebuie respectate pentru a realiza modelarea econometrică sunt:

1. variabilele independente X_1 şi X_2 sunt nestocastice;
2. $E(\varepsilon_i) = 0$ sau $E(\varepsilon_i / X_1, X_2) = 0$;
3. homoscedasticitatea erorilor, $V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$;
4. normalitatea erorilor, $\varepsilon_i \sim N(0, \sigma^2)$;
5. necorelarea erorilor, $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$;
6. lipsa corelaţiei dintre variabilele independente şi variabila eroare, $\text{cov}(\varepsilon, X_1) = \text{cov}(\varepsilon, X_2) = 0$;
7. între variabilele independente nu există o legătură liniară.

Elementele metodologice privind testarea acestor ipoteze sunt prezentate în capitolul 7.

4.2. Estimarea parametrilor modelului

Considerăm modelul de regresie liniară multiplă, cu două variabile independente, la nivelul unei populaţii:

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ sau $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

La nivelul unui eşantion, acesta devine:

$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\varepsilon}_i$ sau

$y_i = \hat{y}_i + \hat{\varepsilon}_i$,

unde $\hat{y}_i$ estimează media condiţionată $E(Y / X_j)$.

Rezultă relaţiile:

$\hat{\varepsilon}_i = y_i - \hat{y}_i$ sau

$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}$.

1. Estimarea parametrilor prin metoda celor mai mici pătrate

Potrivit acestei metode, estimatorii parametrilor modelului de regresie respectă condiția:

$$\sum_i \hat{\varepsilon}_i^2 = \min \text{ sau}$$

$$\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 = \min .$$

Prin rezolvarea problemei de minim, se obține sistemul de ecuații:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_{1i} + \hat{\beta}_2 \sum_i x_{2i} = \sum_i y_i \\ \hat{\beta}_0 \sum_i x_{1i} + \hat{\beta}_1 \sum_i x_{1i}^2 + \hat{\beta}_2 \sum_i x_{1i} x_{2i} = \sum_i y_i x_{1i} \\ \hat{\beta}_0 \sum_i x_{2i} + \hat{\beta}_1 \sum_i x_{1i} x_{2i} + \hat{\beta}_2 \sum_i x_{2i}^2 = \sum_i y_i x_{2i} \end{cases}$$

Rezolvând sistemul, se obțin relațiile pentru estimatorii parametrilor modelului:

$$\hat{\beta}_0 = \bar{\hat{y}} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

$$\hat{\beta}_1 = \frac{\left[\sum_i (y_i - \hat{\bar{y}})(x_{1i} - \bar{x}_1)\right]\sum_i (x_{2i} - \bar{x}_2)^2 - \left[\sum_i (y_i - \hat{\bar{y}})(x_{2i} - \bar{x}_2)\right]\cdot\left[\sum_i (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right]}{\left[\sum_i (x_{1i} - \bar{x}_1)^2\right]\cdot\left[\sum_i (x_{2i} - \bar{x}_2)^2\right] - \left[\sum_i (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right]^2}$$

$$\hat{\beta}_2 = \frac{\left[\sum_i (y_i - \hat{\bar{y}})(x_{2i} - \bar{x}_2)\right]\sum_i (x_{1i} - \bar{x}_1)^2 - \left[\sum_i (y_i - \hat{\bar{y}})(x_{1i} - \bar{x}_1)\right]\cdot\left[\sum_i (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right]}{\left[\sum_i (x_{1i} - \bar{x}_1)^2\right]\cdot\left[\sum_i (x_{2i} - \bar{x}_2)^2\right] - \left[\sum_i (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right]^2}$$

Rezolvarea ecuației matriceale

Sistemul de ecuații rezultat din metoda celor mai mici pătrate se poate scrie ca o ecuație matriceală de forma:

$X' \cdot X \cdot \hat{\beta} = X' \cdot Y$, unde X' este matricea transpusă a matricei X.

Soluția ecuației matriceale ce oferă estimatorii parametrilor este:

$\hat{\beta} = (X' \cdot X)^{-1} \cdot X' \cdot Y$.

BCU "MIHAIL KOGĂLNICEANU" UNIVERSITY LIBRARY

2. Proprietăţile estimatorilor $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

Proprietăţile estimatorilor parametrilor modelului de regresie se pot determina pe baza relaţiei:

$$\hat{\beta} = (X' \cdot X)^{-1} X' \cdot Y = (X' \cdot X)^{-1} X' (X \cdot \beta + \varepsilon) = \beta + (X' \cdot X)^{-1} X' \cdot \varepsilon .$$

Pentru ecuaţia de mai sus, sunt valabile proprietăţile:

- nedeplasare: $E(\hat{\beta}) = \beta$;
- matricea varianţei estimatorilor este $\Omega_{\hat{\beta}} = \sigma^2 (X' \cdot X)^{-1}$, unde:

$\varepsilon_i \sim N(0, \sigma^2)$, iar un estimator al dispersiei erorilor este dat prin relaţia

$$\hat{\sigma}^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n-3} = \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2}{n-3};$$

- estimatorii sunt normal distribuiţi: $\hat{\beta}_j \sim N(\beta_j, \sigma^2_{\hat{\beta}_j})$, iar

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim N(0,1) \text{ şi}$$

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t(n-3).$$

Dacă notăm prin $R_{12}^2 = \dfrac{\left[\sum_i (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right]^2}{\left[\sum_i (x_{1i} - \bar{x}_1)^2\right] \cdot \left[\sum_i (x_{2i} - \bar{x}_2)^2\right]}$

raportul de determinaţie dintre variabilele independente, atunci au loc relaţiile:

$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_{1i} - \bar{x}_1)^2 (1 - R_{12}^2)};$$

$$V(\hat{\beta}_2) = \frac{\sigma^2}{\sum_i (x_{2i} - \bar{x}_2)^2 (1 - R_{12}^2)};$$

BCU IAŞI CENTRAL UNIVERSITY LIBRARY

$$\operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2)=\frac{-R_{12}\sigma^2}{\sqrt{\left[\sum_i (x_{1i}-\bar{x}_1)^2\right]\cdot\left[\sum_i (x_{2i}-\bar{x}_2)^2\right]}\cdot(1-R_{12}^2)}.$$

Dacă în relația $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{1i}+\hat{\beta}_2x_{2i}$ se înlocuiește

$\hat{\beta}_0=\hat{\bar{y}}-\hat{\beta}_1\bar{x}_1-\hat{\beta}_2\bar{x}_2$,

atunci rezultă:

$\hat{y}_i-\hat{\bar{y}}=\hat{\beta}_1(x_{1i}-\bar{x}_1)+\hat{\beta}_2(x_{2i}-\bar{x}_2)$.

Cum $y_i=\hat{y}_i+\hat{\varepsilon}_i$, rezultă proprietatea:

$y_i-\hat{\bar{y}}=\hat{\beta}_1(x_{1i}-\bar{x}_1)+\hat{\beta}_2(x_{2i}-\bar{x}_2)+\hat{\varepsilon}_i$ sau

$\hat{\varepsilon}_i=(y_i-\hat{\bar{y}})-\hat{\beta}_1(x_{1i}-\bar{x}_1)-\hat{\beta}_2(x_{2i}-\bar{x}_2)$.

Din ultima proprietate, se deduce că $\sum_i \hat{\varepsilon}_i=0$.

3. Estimarea punctuală şi prin interval de încredere a parametrilor modelului

Conform proprietății de nedeplasare, pe baza relațiilor estimatorilor, la nivelul unui eşantion de date, se calculează estimații punctuale ale parametrilor modelului.

Pentru un model cu două variabile independente, estimațiile sunt notate cu b_j, $j=\overline{0,2}$.

De exemplu, pentru parametrul β_1, se obține o estimație punctuală prin relația:

$$b_1=\frac{\left[\sum_i (y_i-\bar{y})(x_{1i}-\bar{x}_1)\right]\sum_i (x_{2i}-\bar{x}_2)^2-\left[\sum_i (y_i-\bar{y})(x_{2i}-\bar{x}_2)\right]\cdot\left[\sum_i (x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)\right]}{\left[\sum_i (x_{1i}-\bar{x}_1)^2\right]\cdot\left[\sum_i (x_{2i}-\bar{x}_2)^2\right]-\left[\sum_i (x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)\right]^2}.$$

Intervalele de încredere pentru parametrii modelului de regresie se obțin în baza proprietății de normalitate a estimatorilor. Intervalele de încredere sunt de forma:

$(\hat{\beta}_j\pm t_{\alpha/2}\hat{\sigma}_{\hat{\beta}_j})$, $j=\overline{0,2}$.

Estimatorul varianței estimatorului parametrului β_1 are următoarea relație:

BCU IASI - CENTRAL UNIVERSITY LIBRARY

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum_i (x_{1i} - \bar{x}_1)^2 (1 - R^2_{12})}, \text{ iar}$$

$$\hat{\sigma}^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n-3} = \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2}{n-3}.$$

La nivelul unui eşantion, se obţine un interval estimat, care are relaţia:

$(b_j \pm t_{\alpha/2} s_{\hat{\beta}_j})$, $j = \overline{0,2}$.

Estimaţia varianţei estimatorului parametrului β_1 are relaţia:

$$s^2_{\hat{\beta}_1} = \frac{s^2}{\sum_i (x_{1i} - \bar{x}_1)^2 (1 - R^2_{12})}, \text{ iar}$$

$$s^2 = \frac{\sum_i e_i^2}{n-3} = \frac{\sum_i (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2}{n-3}.$$

Exemplul 1

La nivelul judeţelor României, pentru variabilele *Rata brută de natalitate* (Y, născuţi vii la 1000 de locuitori), *Rata brută de nupţialitate* (X_1, căsătorii la 1000 de locuitori) şi *Rata brută de divorţialitate* (X_2, divorţuri la 1000 de locuitori), s-a construit un model de regresie multiplă pe baza datelor din anul 2015. Utilizând R, comanda şi rezultatele sunt prezentate mai jos:

```
lm(Nat~Cas+Div)

Call:
lm(formula = Nat ~ Cas + Div)
Coefficients:
(Intercept)      Cas      Div
     6,6019   0,6224  -0,8646
```

Ecuaţia modelului estimat este următoarea:

$Y_X = b_0 + b_1 X_1 + b_2 X_2 = 6,6 + 0,62X_1 - 0,86X_2$.

Valoarea $b_0 = 6,6$ născuţi vii la 1000 de locuitori indică valoarea medie estimată a ratei brute de natalitate pentru un judeţ care, ipotetic, ar înregistra o nupţialitate şi o divorţialitate egale cu zero.

Estimaţia $b_1 = 0,62$ născuţi vii la 1000 de locuitori reprezintă valoarea cu care creşte, în medie, rata brută de natalitate la o creştere absolută cu o unitate a ratei brute de nupţialitate, în condiţiile în care divorţialitatea rămâne constantă.

BCU "CAROL I" UNIVERSITY LIBRARY

Estimația $b_2 = -0,86$ născuți vii la 1000 de locuitori reprezintă valoarea cu care scade, în medie, rata brută de natalitate la o creştere cu o unitate a ratei brute de divorțialitate, în condițiile în care nivelul nupțialității rămâne constant.

Din rezultatele de mai sus, se poate observa că nupțialitatea are un impact pozitiv asupra natalității, în timp ce divorțialitatea are un impact negativ.

Pentru calculul intervalelor de încredere, utilizăm comanda:

```
> confint(lm(Nat~Cas+Div))
```

	2,5%	97,5%
(Intercept)	4,7470551	8,45664653
Cas	0,2945087	0,95027371
Div	-1,7743149	0,04520708

4.3. Indicatori de corelație

Pentru un model de regresie liniară multiplă, se pot determina următorii coeficienți: coeficienți de corelație simplă între variabila dependentă şi fiecare variabilă independentă (coeficienți bivariați), coeficienți de corelație parțială (pentru fiecare variabilă independentă în parte), coeficientul de corelație multiplă şi coeficientul de determinație multiplă.

4.3.1. Coeficienți de corelație bivariată, parțială şi semi-parțială

Pentru un model de regresie liniară multiplă cu două variabile independente, se pot determina următorii indicatori de corelație: bivariați, parțiali şi semi-parțiali. Prezentăm ecuațiile estimațiilor acestor indicatori de corelație.

Coeficienți de corelație bivariată

Notăm cu r_{y1}, r_{y2}, r_{12} estimațiile pentru *coeficienții de corelație bivariată* (*zero order*), care măsoară intensitatea legăturii perechilor de două variabile din cele care compun modelul. Ecuațiile sunt:

$$r_{y1} = \frac{n\sum_i x_{1i}y_i - \sum_i x_{1i}\sum_i y_i}{\sqrt{\left[n\sum_i x_{1i}^2 - (\sum_i x_{1i})^2\right]\left[n\sum_i y_i^2 - (\sum_i y_i)^2\right]}};$$

$$r_{y2} = \frac{n\sum_i x_{2i}y_i - \sum_i x_{2i}\sum_i y_i}{\sqrt{\left[n\sum_i x_{2i}^2 - (\sum_i x_{2i})^2\right]\left[n\sum_i y_i^2 - (\sum_i y_i)^2\right]}};$$

$$r_{12}=\frac{n\sum_i x_{1i}x_{2i}-\sum_i x_{1i}\sum_i x_{2i}}{\sqrt{\left[n\sum_i x_{1i}^2-(\sum_i x_{1i})^2\right]\left[n\sum_i x_{2i}^2-(\sum_i x_{2i})^2\right]}}.$$

Observaţie

Se pot scrie relaţiile estimaţiilor coeficienţilor de regresie pe baza coeficienţilor de corelaţie bivariată şi a estimaţiilor abaterilor standard ale variabilelor din model (s'_y, s'_1, s'_2) după cum urmează:

$$b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2;$$

$$b_1=\frac{\sum_i(y_i-\bar{y})(x_{1i}-\bar{x}_1)-r_{12}r_{y2}s'_y s'_1}{s'^2_1(1-r^2_{12})} \text{ sau } b_1=\frac{s'_y}{s'_1}\frac{r_{y1}-r_{y2}r_{12}}{(1-r^2_{12})};$$

$$b_2=\frac{\sum_i(y_i-\bar{y})(x_{2i}-\bar{x}_2)-r_{12}r_{y1}s'_y s'_2}{s'^2_2(1-r^2_{12})} \text{ sau } b_2=\frac{s'_y}{s'_2}\frac{r_{y2}-r_{y1}r_{12}}{(1-r^2_{12})}.$$

Coeficienţi de corelaţie parţială

La nivelul unui eşantion, estimaţiile pentru *coeficienţii de corelaţie parţială* se vor nota prin:

- $r_{y1.2}$, coeficientul de corelaţie dintre Y şi X_1, în condiţiile în care influenţa variabilei X_2 este considerată constantă;
- $r_{y2.1}$, coeficientul de corelaţie dintre Y şi X_2, în condiţiile în care influenţa variabilei X_1 este considerată constantă;

Pentru fiecare coeficient de corelaţie parţială, estimaţiile sunt date prin:

$$r_{y1.2}=\frac{r_{y1}-r_{y2}r_{12}}{\sqrt{(1-r^2_{y2})(1-r^2_{12})}};\ r_{y2.1}=\frac{r_{y2}-r_{y1}r_{12}}{\sqrt{(1-r^2_{y1})(1-r^2_{12})}}.$$

Observaţia 1

Coeficientul $r^2_{y1.2}$ se numeşte *coeficient de determinaţie parţială* şi arată cât la sută din variaţia lui Y a rămas neexplicată din modelul dintre Y şi X_2 şi este explicată prin introducerea în model a variabilei X_1.

Observaţia 2

Aşa cum se observă din relaţiile de mai sus, în evaluarea impactului unei anumite variabile independente asupra celei dependente se cuantifică şi o componentă care este determinată de efectul unei variabile independente asupra

celeilalte (r_{12}). Cu alte cuvinte, pentru un coeficient de corelație parțială se menține constantă influența unei variabile independente, atât asupra celei dependente, cât și asupra celeilalte variabile independente.

Coeficienți de corelație semi-parțială

În cazul acestor coeficienți, se consideră constantă influența variabilei independente doar asupra variabilei dependente, eliminând efectul variației comune existente între variabilele independente. Cu alte cuvinte, un asemenea coeficient arată contribuția unică a unei variabile independente asupra celei dependente.

Relațiile celor doi *coeficienți de corelație semi-parțială* sunt:

$$r_{y(1.2)} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{12}^2)}};\ r_{y(2.1)} = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{(1-r_{12}^2)}}.$$

Importanța acestor coeficienți poate fi subliniată din următoarele relații:

$R^2 = r_{y1}^2 + r_{y(2.1)}^2$ sau $r_{y(2.1)}^2 = R^2 - r_{y1}^2$;

$R^2 = r_{y2}^2 + r_{y(1.2)}^2$ sau $r_{y(1.2)}^2 = R^2 - r_{y2}^2$.

$r_{y(1.2)}^2$ este un *coeficient de determinație semi-parțială* și arată cu cât se modifică coeficientul de determinație (R^2) dacă variabila X_1 ar fi introdusă în sau exclusă din modelul de regresie. Au loc și relațiile:

$$r_{y1.2} = \frac{r_{y(1.2)}}{\sqrt{(1-r_{y2}^2)}} \text{ sau } r_{y(1.2)}^2 = r_{y1.2}^2(1-r_{y2}^2).$$

$$r_{y2.1} = \frac{r_{y(2.1)}}{\sqrt{(1-r_{y1}^2)}} \text{ sau } r_{y(2.1)}^2 = r_{y2.1}^2(1-r_{y1}^2).$$

Exemplul 2

Utilizăm variabilele de la exemplul 1 și estimăm coeficienții de corelație bivariată și parțială cu ajutorul programului R. Pentru corelația bivariată, se aplică funcția *cor* pentru o structură de date (obținută cu funcția *data.frame*) care conține variabilele : *Nat*, *Cas*, *Div*. Rezultă matricea de corelație de mai jos.

	Nat	Cas	Div
Nat	1,0000000	0,4829942	-0,1868171
Cas	0,4829942	1,0000000	0,1411820
Div	-0,1868171	0,1411820	1,0000000

Coeficientul de corelație bivariată între natalitate și nupțialitate este $r_{y1} = 0{,}48$. Acesta arată că între cele două variabile există o legătură liniară directă

şi de intensitate medie. Al doilea coeficient de corelaţie bivariată este $r_{y2} = -0{,}18$ şi indică o legătură inversă slabă între natalitate şi divorţialitate.

Coeficienţii de corelaţie parţială se obţin cu ajutorul funcţiei *pcor* (funcţia este disponibilă în pachetul *ppcor*) aplicate structurii de date. În R, rezultatele sunt:

	Nat	Cas	Div
Nat	1,0000000	0,5237437	-0,2941758
Cas	0,5237437	1,0000000	0,2690205
Div	-0,2941758	0,2690205	1,0000000

În tabelul de rezultate, sunt prezenţi ambii coeficienţi de corelaţie parţială, pe linia sau coloana variabilei dependente:

- $r_{y1.2} = 0{,}52$ şi indică o influenţă parţială pozitivă, de intensitate medie, a nupţialităţii asupra natalităţii; $r^2_{y1.2} = 27\%$ arată că 27% din variaţia natalităţii neexplicată de legătura dintre natalitate şi divorţialitate este explicată prin adăugarea în model a variabilei nupţialitate;
- $r_{y2.1} = -0{,}29$ şi indică o influenţă parţială negativă, de intensitate slabă, a divorţialităţii asupra natalităţii.

Coeficienţii de corelaţie semi-parţială se obţin cu ajutorul funcţiei *spcor* aplicate structurii de date ce este compusă din variabilele supuse analizei. Pentru cele trei variabile, rezultatele obţinute în R sunt: $r_{y(1.2)} = 0{,}51$; $r_{y(2.1)} = -0{,}26$.

Valoarea $r^2_{y(2.1)} = 6{,}7\%$ arată că, prin eliminarea variabilei *rata divorţialităţii* din modelul iniţial, nivelul coeficientului de determinaţie (pentru modelul fără această variabilă) se reduce cu 0,067.

4.3.2. Raportul de determinaţie multiplă şi raportul de corelaţie multiplă

Raportul de determinaţie multiplă sau *coeficientul de determinaţie multiplă* arată ponderea din variaţia totală a variabilei dependente care este explicată de variaţia simultană a variabilelor independente incluse în model.

Relaţia raportului de determinaţie multiplă este:

$$\eta^2 = \frac{\sum_i (y_{x_i} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}.$$

Raportul de corelaţie multiplă măsoară gradul de intensitate a legăturii simultane dintre variabila dependentă şi variabilele independente şi are relaţia:

$$\eta = \sqrt{\eta^2} = \sqrt{\frac{\sum_i (y_{x_i} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}}, \text{ iar } 0 < \eta < 1.$$

Estimatorul raportului de determinaţie multiplă

Pentru un model liniar multiplu cu două variabile independente, estimatorul raportului de determinaţie multiplă se determină pe baza relaţiei de descompunere a estimatorului variaţiei totale:

$$\hat{V}_T = \hat{V}_E + \hat{V}_R \quad \text{sau}$$

$$\sum_i (y_i - \hat{\bar{y}})^2 = \sum_i (\hat{y}_i - \hat{\bar{y}})^2 + \sum_i (y_i - \hat{y}_i)^2, \text{ iar}$$

$$\hat{\eta}^2 = \frac{\hat{V}_E}{\hat{V}_T} = \frac{\hat{\beta}_1 \sum_i (y_i - \hat{\bar{y}})(x_{1i} - \bar{x}_1) + \hat{\beta}_2 \sum_i (y_i - \hat{\bar{y}})(x_{2i} - \bar{x}_2)}{\sum_i (y_i - \hat{\bar{y}})^2}.$$

Relaţia de mai sus se poate scrie astfel:

$$\hat{\eta}^2 = 1 - \frac{\hat{V}_R}{\hat{V}_T} = 1 - \frac{\sum_i \hat{\varepsilon}_i^2}{\sum_i (y_i - \hat{\bar{y}})^2}.$$

O estimaţie a raportului sau coeficientului de determinaţie multiplă este:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}.$$

Exemplul 3

Considerăm variabilele prezentate la exemplul 1. Pentru modelul de regresie multiplă, în R, se poate obține un set de rezultate cu ajutorul funcției *summary*.

Astfel, se aplică comanda:

```
> summary(lm(Nat~Cas+Div))
```

La finalul setului de rezultate, se prezintă estimația raportului de determinație:

Multiple R-squared: 0,2996

Această valoare se poate calcula cu ajutorul estimațiilor componentelor celor două dispersii, explicată și reziduală, care se obțin cu ajutorul *analizei varianței* (*ANOVA*). În R, aceste rezultate se obțin aplicând funcția *anova* obiectului obținut din analiza de regresie. Comanda și rezultatele sunt prezentate mai jos:

BCU IASI / CENTRAL UNIVERSITY LIBRARY

```
> anova(lm(Nat~Cas+Div))
Analysis of Variance Table
Response: Nat
          Df  Sum Sq  Mean Sq  F value     Pr(>F)
Cas        1  10,684  10,6842  12,9904  0,0008759 ***
Div        1   3,039   3,0388   3,6948  0,0619100 .
Residuals 39  32,076   0,8225
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

În R, componentele variației sunt prezentate pe fiecare variabilă independentă și pe componenta reziduală. Pentru a obține variația explicată (*ESS*), este nevoie să se însumeze variațiile (*Sum Sq*) pentru cele două variabile din model. Variația totală (*TSS*) este suma tuturor variațiilor (din coloana *Sum Sq*).

Pe baza tabelului de rezultate, se poate determina estimația raportului de determinație multiplă:

$$R^2 = \frac{ESS}{TSS} = \frac{10,684+3,039}{10,684+3,039+32,076} = 0,299.$$

Interpretare

Variația natalității la nivelul județelor este explicată în proporție de 29,9% de variația simultană a nupțialității și divorțialității. Restul de până la 100% reprezintă influența altor factori neincluși în model.

4.3.3. Coeficientul de corelație multiplă

Pe baza coeficienților de corelație bivariată și a celor parțiali, se pot obține o serie de relații pentru estimația coeficientului de corelație multiplă (r):

$$r = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1-r_{12}^2}} \text{ sau}$$

$$r = \sqrt{r_{y1}^2 + (1-r_{y1}^2)r_{y.2.1}} \text{ sau } r = \sqrt{r_{y2}^2 + (1-r_{y2}^2)r_{y1.2}}.$$

Coeficientul de corelație multiplă este un indicator care măsoară intensitatea legăturii dintre variabila dependentă și toate variabilele independente cuprinse în model.

Exemplul 4

Utilizăm matricea de corelație determinată la exemplul 2, pentru cele 3 variabile analizate: natalitate, nupțialitate și divorțialitate.

Cei 3 coeficienți de corelație bivariați sunt:

$r_{y1} = 0,482$; $r_{y2} = -0,186$; $r_{12} = 0,141$.

BCU "ION CREANGĂ" UNIVERSITY LIBRARY

Așadar,

$$r = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}} = 0{,}52 .$$

Acest coeficient arată că între cele trei variabile există o legătură liniară de intensitate medie.

4.3.4. Raportul de determinație multiplă ajustat

Raportul de determinație multiplă nu ține seama de numărul de grade de libertate sau de numărul de parametri care apar în model. În consecință, în evaluarea intensității legăturii dintre variabile, se utilizează un *coeficient de determinație ajustat* care ia în calcul acest număr de parametri.

Estimatorul coeficientului de determinație ajustat are următoarea relație:

$$\hat{\bar{\eta}}^2 = 1 - \frac{\frac{\hat{V}_R}{n-k}}{\frac{\hat{V}_T}{n-1}} = 1 - \frac{\frac{\sum_i \hat{\varepsilon}_i^2}{n-k}}{\frac{\sum_i (y_i - \hat{\bar{y}})^2}{n-1}} = 1 - \frac{\hat{\sigma}^2}{V(Y)}.$$

Scris altfel,

$$\hat{\bar{\eta}}^2 = 1 - (1 - \hat{\eta}^2)\frac{n-1}{n-k}.$$

Estimația raportului de determinație ajustat este:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \text{ sau}$$

$$\bar{R}^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}}.$$

Observație

Din relațiile de mai sus, se poate observa că pentru estimațiile celor doi estimatori, are loc relația: $\bar{R}^2 < R^2$, pentru $k > 1$.

Exemplul 5

Funcția *summary* din R, aplicată obiectului obținut prin modelarea econometrică, returnează atât valoarea estimată a raportului de determinație, cât și a raportului de determinație ajustat. Pentru datele prezentate la exemplul 1, rezultatele sunt următoarele:

BCU IAȘI CENTRAL UNIVERSITY LIBRARY

Multiple R-squared: 0,299, Adjusted R-squared: 0,264

Pe baza datelor din tabelul *ANOVA* obținute la exemplul 3, se cunosc:

$RSS = 32,076$;

$TSS = 45,799$.

De asemenea, se cunosc datele: $n = 42$, $k = 3$ (numărul de parametri din model). Rezultă:

$$\overline{R}^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} = 1 - \frac{\frac{32,076}{42-3}}{\frac{45,799}{42-1}} = 0,264.$$

Observație

Rezultatul obținut are aceeași interpretare cu valoarea $R^2 = 0,299$. Se confirmă faptul că $\overline{R}^2 < R^2$, însă valoarea $\overline{R}^2 = 0,264$ este mai precisă.

4.4. Testarea parametrilor și a modelului

Pentru un model de regresie multiplă, se pot construi mai multe teste, cu scopul de a testa: parametrii modelului, modelul de regresie, influența marginală a unei variabile etc.

4.4.1. Testarea coeficienților de regresie

Parametrii modelului de regresie liniară multiplă se testează cu ajutorul testului Student, considerând estimatorii obținuți prin metoda celor mai mici pătrate și legea de repartiție a acestora. Pentru cazul modelului liniar multiplu cu două variabile independente, demersul testării este prezentat mai jos.

1. Formularea ipotezelor

$H_0: \beta_j = 0$, $j = \overline{1,2}$ (variabila independentă X_j nu are o influență liniară parțială asupra celei dependente);

$H_1: \beta_j \neq 0$, $j = \overline{1,2}$.

2. Alegerea pragului de semnificație α

De regulă, se consideră $\alpha = 0,05$.

3. Alegerea testului

În acest caz, se utilizează statistica Student. În condițiile acceptării ipotezei nule, pentru fiecare parametru, se utilizează statistica:

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}, \quad j = \overline{1,2},$$

care urmează o lege de repartiție Student de $n-k=n-3$ grade de libertate.

4. Valoarea teoretică a statisticii

Pentru pragul de semnificație ales și pentru legea de repartiție cunoscută a testului, se citește valoarea teoretică din tabela Student: $t_{\alpha/2;n-k} = t_{\alpha/2;n-3}$.

5. Valoarea calculată a testului

La nivelul unui eșantion, se determină valoarea calculată a testului:

$$t_{calc} = \frac{b_j}{s_{\hat{\beta}_j}}, \ j = \overline{1,2}.$$

6. Regula de decizie

- dacă $t_{calc} \in [-t_{\alpha/2;n-3}, t_{\alpha/2;n-3}]$, se acceptă ipoteza H_0;
- dacă $t_{calc} \notin [-t_{\alpha/2;n-3}, t_{\alpha/2;n-3}]$, se respinge H_0, cu probabilitatea $(1-\alpha)$.

În programele de statistică, decizia se ia pe baza semnificației testului:

- dacă $Sig\, t < \alpha$, se respinge H_0, cu nivelul de încredere specificat;
- dacă $Sig\, t \geq \alpha$, nu se respinge ipoteza nulă.

Exemplul 6

Considerăm variabilele prezentate la exemplul 1. De asemenea, aplicăm comanda:

```
> summary(lm(Nat~Cas+Div))
```

Rezultatele privitoare la modelul de regresie multiplă se prezintă mai jos:

```
Call:
lm(formula = Nat ~ Cas + Div)
Residuals:
     Min       1Q   Median       3Q      Max
-1,4031  -0,6819  -0,2026   0,8228   1,5215
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6,6019     0,9170   7,199 1,14e-08 ***
Cas           0,6224     0,1621   3,840 0,000441 ***
Div          -0,8646     0,4498  -1,922 0,061910 .
---
Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

BCU "CAROL I" CENTRAL UNIVERSITY LIBRARY

Rezultatele din tabelul de mai sus permit testarea parametrilor modelului de regresie cu ajutorul valorii semnificației testului (*Sig t*), care este prezentată în coloana $Pr(>|t|)$. Se observă că, pentru primii doi parametri din model, se respectă condiția $Sigt < 0,05$. În concluzie, cu o probabilitate de 0,95, se consideră că parametrii β_0, β_1 sunt semnificativi statistic.

Așadar, putem considera că variabila nupțialitate are un impact semnificativ asupra natalității, cu o probabilitate de 0,95.

Pentru parametrul β_2, nu există motive de a respinge ipoteza nulă, decât dacă se consideră un nivel de încredere mai scăzut, cum ar fi cel de 0,9.

Dacă se dorește testarea cu demersul clasic pe baza valorilor teoretice și calculate ale testului, se utilizează coloana *t value* din tabelul de mai sus (care conține valorile calculate ale testului pentru fiecare parametru) și se citește din tabelul Student valoarea corespunzătoare pentru un prag de semnificație de 0,05 și $n-3=39$ de grade de libertate (valoarea este egală cu 2,02). Aplicând regula de decizie, se obține același rezultat.

În cazul rezultatelor din R, valoarea calculată a testului Student pentru un parametru se poate obține ca raport între estimația parametrului și estimația erorii standard corespunzătoare:

$$t_{calc} = \frac{Estimate}{Std.Error}.$$

4.4.2. Testarea modelului de regresie

Testarea modelului de regresie liniară multiplă se realizează cu ajutorul testului *F*, conform următorului demers:

1. Formularea ipotezelor

$H_0: \beta_0 = \beta_1 = ... = \beta_p = 0$ (modelul nu este semnificativ);

H_1: nu toți coeficienții variabilelor independente sunt simultan zero (există un model de regresie semnificativ).

2. Alegerea pragului de semnificație α

3. Alegerea testului

Se utilizează statistica Fisher de forma:

$$F = \frac{\hat{V}_E}{\hat{V}_R} \cdot \frac{n-k}{k-1} = \frac{\hat{\eta}^2}{1-\hat{\eta}^2} \cdot \frac{n-k}{k-1} \sim F(k-1, n-k).$$

Observație

Pentru două variabile independente, $k=3$, statistica Fisher este:

$$F = \frac{\hat{V}_E}{\hat{V}_R} \cdot \frac{n-k}{k-1} = \frac{\hat{V}_E}{\hat{V}_R} \cdot \frac{n-3}{2} \text{ sau}$$

$$F = \frac{\hat{\beta}_1 \sum_i (y_i - \bar{\hat{y}})(x_{1i} - \bar{x}_1) + \hat{\beta}_2 \sum_i (y_i - \bar{\hat{y}})(x_{2i} - \bar{x}_2)}{\sum_i \hat{\varepsilon}_i^2} \cdot \frac{n-3}{2}.$$

Dacă se consideră estimatorul coeficientului de determinație, statistica Fisher devine:

$$F = \frac{\hat{\eta}^2}{1-\hat{\eta}^2} \cdot \frac{n-k}{k-1} = \frac{\hat{\eta}^2}{1-\hat{\eta}^2} \cdot \frac{n-3}{2} \sim F(2, n-3).$$

4. Valoarea teoretică a statisticii

Se citește valoarea teoretică din tabela Fisher: $F_{\alpha;k-1;n-k}$.

5. Valoarea calculată a testului

La nivelul unui eșantion, se obține valoarea calculată a testului:

$$F_{calc} = \frac{ESS}{RSS} \cdot \frac{n-k}{k-1} = \frac{R^2}{1-R^2} \cdot \frac{n-k}{k-1}.$$

6. Regula de decizie

- dacă $F_{calc} > F_{\alpha;k-1;n-k}$, se respinge ipoteza H_0, cu probabilitatea $(1-\alpha)$;
- dacă $F_{calc} \leq F_{\alpha;k-1;n-k}$, se acceptă ipoteza nulă.

Pe baza semnificației testului, dacă $SigF < \alpha$, se respinge H_0, cu probabilitatea $(1-\alpha)$. În caz contrar, nu se respinge ipoteza nulă.

Exemplul 7

Considerăm variabilele prezentate la exemplul 1. Pentru a obține rezultatele privind coeficientul de determinație și testul Fisher aferent, aplicăm comanda:

```
> summary(lm(Nat~Cas+Div))
```

La finalul setului de rezultate obținute în R, se oferă informații despre statistica Fisher care testează semnificația raportului de determinație sau semnificația modelului de regresie. Rezultatele sunt următoarele:

```
Residual standard error: 0,9069 on 39 degrees of freedom
Multiple R-squared: 0,2996,      Adjusted R-squared: 0,2637
F-statistic: 8,343 on 2 and 39 DF,  p-value: 0,0009635
```

Valoarea calculată *F-statistic* se obține cu ajutorul relației:

BCU "CAROL I" CENTRAL UNIVERSITY LIBRARY

$$F_{calc} = \frac{R^2}{1-R^2} \cdot \frac{n-k}{k-1} = \frac{0,2996}{0,7004} \cdot \frac{39}{2} = 8,34.$$

Semnificația testului (*p-value*) este 0,00096 și este mai mică decât pragul de semnificație 0,05. În concluzie, se respinge ipoteza nulă, adică se poate considera că modelul de regresie explică semnificativ legătura dintre variabile.

Același rezultat se obține dacă se consideră componentele variației, așa cum au fost prezentate la exemplul 3:

$ESS = 13,723$;

$RSS = 32,076$.

Știind că $n = 42$, $k = 3$, se obține:

$$F_{calc} = \frac{ESS}{RSS} \cdot \frac{n-k}{k-1} = \frac{13,723}{32,076} \cdot \frac{39}{2} = 8,34.$$

Din tabela Fisher, pentru un risc de 0,05, $k - 1 = 2$ și $n - k = 39$ de grade de libertate, se citește valoarea $F_{0,05;2;39} = 3,232$.

Decizia

Deoarece $F_{calc} > F_{0,05;2;39}$, se respinge ipoteza nulă, cu o probabilitate de 0,95, adică modelul explică semnificativ dependența dintre variabilele utilizate.

Observație

În R, componentele variației și testul Fisher sunt determinate pentru fiecare variabilă independentă separat:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cas	1	10,684	10,6842	12,9904	0,0008759
Div	1	3,039	3,0388	3,6948	0,0619100
Residuals	39	32,076	0,8225		

Fiecare test Fisher are un corespondent în testul Student ce verifică semnificația parametrului de regresie corespunzător.

4.4.3. Testarea influenței marginale a unei variabile independente

În cazul în care se dorește testarea influenței marginale a unei variabile nou introduse în model sau a uneia excluse din model, se folosește un test Fisher, cu o statistică dată prin relația:

$$F = \frac{\hat{V}_{E\_new} - \hat{V}_{E\_old}}{\hat{V}_{R\_new}} \cdot \frac{n-k_{new}}{k_{new}-k_{old}} = \frac{\hat{\eta}^2_{new} - \hat{\eta}^2_{old}}{1-\hat{\eta}^2_{new}} \cdot \frac{n-k_{new}}{k_{new}-k_{old}},$$

unde „old" specifică indicatorul înainte de introducerea în model a variabilei independente, iar „new", indicatorul după introducerea variabilei în model.

În acest caz, ipotezele sunt următoarele:

H_0: variabila introdusă în model nu are o influență semnificativă asupra variației variabilei dependente;

H_1: variabila are o influență semnificativă asupra variației variabilei dependente.

Valoarea calculată a testului este:

$$F_{calc} = \frac{ESS_{new} - ESS_{old}}{RSS_{new}} \cdot \frac{n - k_{new}}{k_{new} - k_{old}} = \frac{R^2_{new} - R^2_{old}}{1 - R^2_{new}} \cdot \frac{n - k_{new}}{k_{new} - k_{old}}.$$

Decizia de a accepta sau a respinge ipoteza nulă se ia în aceleași condiții precizate mai sus pentru testul Fisher.

Observație

Dacă se consideră situația excluderii variabilelor nesemnificative din model, atunci în relația de mai sus se schimbă poziția „old" cu „new" pentru raportul de determinație, precum și pentru numărul de grade de libertate, astfel:

$$F_{calc} = \frac{R^2_{old} - R^2_{new}}{1 - R^2_{old}} \cdot \frac{n - k_{old}}{k_{old} - k_{new}} \text{ sau}$$

$$F_{calc} = \frac{RSS_{new} - RSS_{old}}{RSS_{old}} \cdot \frac{n - k_{old}}{(n - k_{new}) - (n - k_{old})}.$$

Exemplul 8

Pentru județele României, în anul 2015, considerăm dependența dintre natalitate (variabila dependentă) și nupțialitate, divorțialitate, mortalitate infantilă și mortalitate (variabile independente).

În R, rezultatele modelării se obțin astfel:

```
> summary(lm(Nat~Cas+Div+Inf))

Call:
lm(formula = Nat ~ Cas + Div + Inf)
Residuals:
     Min       1Q   Median       3Q      Max
-1,4496  -0,6793  -0,1306   0,7866   1,6361
Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  5,82016   1,54862    3,758    0,000574 ***
Cas          0,69680   0,20170    3,455    0,001370 **
Div         -0,90241   0,45728   -1,973    0,055751 .
Inf          0,05092   0,08094    0,629    0,533069
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

Pentru testarea influenței marginale a variabilelor independente, în R, se utilizează metoda *Backward*. Această metodă exclude pe rând variabilele independente nesemnificative din modelul liniar inițial care conține toate variabilele. Să notăm modelul inițial *model1*. Acesta se obține în R astfel:

```
> model1=lm(Nat~Cas+Div+Inf)
```

Prin metoda *Backward*, R exclude în prima etapă variabila cea mai nesemnificativă din model. Se obține un nou model cu o variabilă mai puțin. Notăm acest nou model cu *model2*. În R, acesta se obține astfel:

```
> model2=step(model1, direction="backward")
```

```
Start:  AIC = -3,76
Nat ~ Cas + Div + Inf
        Df   Sum Sq   RSS     AIC
- Inf    1   0,3306  32,076  -5,3215
<none>              31,745  -3,7566
- Div    1   3,2534  34,999  -1,6588
- Cas    1   9,9701  41,716   5,7146
Step:  AIC = -5,32
Nat ~ Cas + Div
        Df   Sum Sq   RSS     AIC
<none>              32,076  -5,3215
- Div    1   3,0388  35,115  -3,5199
- Cas    1  12,1246  44,201   6,1449
```

Dacă analizăm rezultatele de mai sus, observăm că pentru primul model variația reziduală are valoarea $RSS_{old} = 31,745$. De asemenea, criteriul informațional asociat este $AIC = -3,76$.

Metoda *Backward* presupune construirea tuturor modelelor cu 2 variabile independente, prin excluderea pe rând a câte unei variabile independente. Modelul cu 2 variabile cel mai bun se alege cu ajutorul criteriului informațional *AIC*. Se observă că acest model nu conține variabila mortalitate infantilă *(Inf)*, iar valoarea criteriului informațional este: $AIC = -5,32$.

Testarea influenței marginale a variabilei excluse *(Inf)* se realizează cu ajutorul comenzii:

```
> anova(model2,model1)
Analysis of Variance Table
Model 1: Nat ~ Cas + Div
Model 2: Nat ~ Cas + Div + Inf
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	39	32,076				
2	38	31,745	1	0,33059	0,3957	0,5331

Valoarea calculată a testului Fisher se obține astfel:

$$F_{calc} = \frac{RSS_{new} - RSS_{old}}{RSS_{old}} \cdot \frac{n - k_{old}}{(n - k_{new}) - (n - k_{old})}.$$

Înlocuind datele din tabelul de mai sus, obținem:

$$F_{calc} = \frac{32,076 - 31,754}{31,745} \cdot \frac{38}{39 - 38} = 0,3957.$$

Decizia

Al doilea model se obține prin eliminarea variabilei independente *Inf* din modelul inițial. Această eliminare nu produce o modificare semnificativă a raportului de determinație, așa cum o confirmă testul Fisher.

Semnificația testului Fisher (*Sig. F Change*), care testează modificarea raportului de determinație, este egală cu *0,53*. Acest rezultat indică decizia de a accepta ipoteza nulă, care susține că între cele două modele nu sunt diferențe semnificative, sau decizia de a elimina din model variabila *Inf*.

Procesul de eliminare se oprește pentru că variabile independente rămase sunt semnificative statistic, în limitele unui risc de 1% care este setat implicit.

4.5. Probleme specifice modelelor de regresie liniară multiplă

În cazul modelelor de regresie liniară multiplă, pot exista situații concrete în care procesul de modelare se realizează cu anumite restricții cu privire la parametri. De asemenea, acest tip de modele permite realizarea unei ierarhii a factorilor de influență în funcție de contribuția fiecăruia la variația variabilei dependente. Asupra acestor două probleme ne vom opri în continuare.

4.5.1. Modele cu restricții

Considerăm modelul de regresie liniară multiplă de forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon .$$

Dacă pentru acest model se impune una sau mai multe restricții asupra parametrilor (de exemplu, $\beta_2 = 1$, adică orice creștere a variabilei X_2 cu o unitate determină o creștere medie a rezultatului tot cu o unitate, păstrând constantă influența celorlalte variabile independente), procesul de modelare econometrică permite evaluarea acestor condiții cu ajutorul unui test Fisher.

BCU CENTRAL UNIVERSITY LIBRARY

Ideea de bază a acestui procedeu este aceea de a construi două modele de regresie. Primul este modelul fără nici o restricţie, adică modelul de bază în care apar toate variabilele (de exemplu, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$).

Al doilea model se obţine prin acceptarea restricţiilor. Dacă, de exemplu, se consideră condiţia $\beta_2 = 1$, atunci modelul iniţial devine:

$$Y = \beta_0 + \beta_1 X_1 + X_2 + ... + \beta_p X_p + \varepsilon .$$

Acest ultim model se poate rescrie astfel:

$$Y - X_2 = \beta_0 + \beta_1 X_1 + \beta_3 X_3 ... + \beta_p X_p + \varepsilon \text{ sau}$$

$$Y^* = \beta_0 + \beta_1 X_1 + \beta_3 X_3 ... + \beta_p X_p + \varepsilon .$$

Ce se poate observa din exemplul de restricţie utilizat mai sus este că noul model de regresie obţinut, numit *model cu restricţii*, este un model transformat al celui de bază în care se reduce numărul de parametri în funcţie de restricţiile formulate. Pentru acest gen de problemă, este important atât numărul de restricţii, cât şi tipul lor. Există restricţii simple, formate pe baza unui singur parametru ($\beta_2 = 1$) şi restricţii compuse, care sunt alcătuite din mai mulţi parametri (de exemplu, $\beta_1 + \beta_2 = 1$).

Testarea unui model cu restricţii implică un demers care presupune estimarea în paralel a celor două modele de regresie, cel fără restricţii şi cel cu restricţii, şi apoi parcurgerea următoarelor etape:

1. Formularea ipotezelor

H_0: ipoteza se referă la restricţiile impuse modelului de regresie (de exemplu, $\beta_1 + \beta_2 = 1$);

H_1: reprezintă alternativa la restricţiile impuse (de exemplu, $\beta_1 + \beta_2 \neq 1$).

2. Alegerea pragului de semnificaţie α

3. Alegerea testului

Se utilizează statistica Fisher care are următoarea relaţie:

$$F = \frac{R\hat{V}_R - U\hat{V}_R}{U\hat{V}_R} \cdot \frac{n-k}{m} \sim F(m; n-k) \text{, unde}$$

- $R\hat{V}_R$ reprezintă estimatorul variaţiei reziduale pentru modelul cu restricţii;
- $U\hat{V}_R$ este estimatorul variaţiei reziduale pentru modelul fără restricţii;
- n este volumul eşantionului;
- k este numărul de parametri din modelul de bază, fără restricţii;

BCU "LUCIAN BLAGA" CENTRAL UNIVERSITY LIBRARY

- m este numărul de restricții, indiferent dacă este vorba de restricții simple sau compuse.

4. Valoarea teoretică a statisticii

Se citește valoarea teoretică din tabela Fisher: $F_{\alpha;m;n-k}$.

5. Valoarea calculată a testului

La nivelul unui eșantion, se obține valoarea calculată a testului:

$$F_{calc} = \frac{RRSS - URSS}{URSS} \cdot \frac{n-k}{m}, \text{ unde}$$

- *RRSS* (*Restricted Residual Sum of Squares*) este estimația variației reziduale pentru modelul cu restricții;
- *URSS* (*Unrestricted Residual Sum of Squares*) reprezintă estimația variației reziduale pentru modelul fără restricții.

Observație

În cazul în care modelul cu restricții și modelul fără restricții au aceeași variabilă dependentă, valoarea calculată a testului se poate obține și cu ajutorul estimațiilor coeficienților de determinație din cele două modele, astfel:

$$F_{calc} = \frac{R_U^2 - R_R^2}{1 - R_U^2} \cdot \frac{n-k}{m}, \text{ unde}$$

- R_U^2 este estimația coeficientului de determinație în cazul modelului de bază;
- R_R^2 reprezintă estimația coeficientului de determinație în cazul modelului cu restricții.

6. Regula de decizie

- dacă $F_{calc} > F_{\alpha;m;n-k}$, se respinge ipoteza H_0, cu probabilitatea $(1-\alpha)$;
- dacă $F_{calc} \leq F_{\alpha;m;n-k}$, nu se respinge ipoteza nulă.

Dacă nu există motive să se respingă ipoteza nulă, atunci se poate considera că restricția sau restricțiile impuse pot fi acceptate sau se poate accepta ca ecuație de regresie modelul cu restricții.

Pe baza semnificației testului, dacă $SigF < \alpha$, se respinge H_0, cu probabilitatea $(1-\alpha)$. În caz contrar, se acceptă ipoteza nulă.

Exemplul 9

La nivelul județelor României, în anul 2015, considerăm legătura dintre variabila dependentă natalitate și variabile independente: nupțialitate, divorțialitate, mortalitate, mortalitate infantilă.

BCU CENTRAL UNIVERSITY LIBRARY

Tendința de evoluție a natalității și mortalității în trecut sugerează că între cele două variabile există o legătură inversă. Dorim să verificăm ipoteza conform căreia, la nivel de județ, o creștere absolută cu o unitate a mortalității ar determina o scădere absolută cu 0,5 unități a natalității, în condițiile în care ceilalți factori rămân constanți.

Pentru acest exemplu, ipoteza nulă care trebuie verificată este următoarea:

$H_0: \beta_3 = -0{,}5$.

Așadar, vorbim despre o singură restricție ($m = 1$).

Modelul fără restricții este:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$,

iar rezultatele modelării sunt prezentate în tabelul de mai jos.

```
> model3=lm(Nat~Cas+Div+Mort+Inf.)
> summary(model3)

Call:
lm(formula = Nat ~ Cas + Div + Mort + Inf)
Residuals:
     Min       1Q   Median       3Q      Max
 -1,7187  -0,4874  -0,1143   0,5763   1,7499
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) 11,44504    2,02102     5,663   1,8e-06 ***
Cas          0,39949    0,19190     2,082   0,044341 *
Div         -1,19649    0,40330    -2,967   0,005248 **
Mort        -0,37392    0,10062    -3,716   0,000666 ***
Inf          0,11838    0,07231     1,637   0,110096
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

Modelul cu restricții este:

$Y^* = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \varepsilon$, unde

$Y^* = Y - 0{,}5X_3$.

Testarea în R a restricției de mai sus se realizează ușor prin comanda (funcția este disponibilă în pachetul *car*):

```
> linearHypothesis(model3,"Mort=-0.5")

Linear hypothesis test
Hypothesis:
Mort = - 0.5
Model 1: restricted model
```

Model 2: Nat ~ Cas + Div + Mort + Inf

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	38	24,098				
2	37	23,117	1	0,98097	1,5701	0,2181

Valoarea calculată a testului Fisher este:

$$F_{calc} = \frac{RRSS - URSS}{URSS} \cdot \frac{n-k}{m} = \frac{24,098 - 23,117}{23,117} \cdot \frac{37}{1} = 1,57 .$$

Din tabelul Fisher, pentru un nivel de încredere de 0,95, se citeşte valoarea teoretică $F_{\alpha;m;n-k} = F_{0,05;1;37} = 4,1$.

Deoarece $F_{calc} \leq F_{\alpha;m;n-k}$, nu avem motive să respingem ipoteza nulă, ceea ce înseamnă că se poate accepta modelul cu restricţii pentru explicarea variaţiei natalităţii.

4.5.2. Modele cu variabile standardizate

Unul dintre obiectivele principale ale modelării econometrice cu ajutorul modelului de regresie liniară multiplă este acela de a stabili, pe lângă factorii care au un impact semnificativ asupra unei variabile dependente, care este ordinea importanţei acestora. Pentru a atinge acest obiectiv, se construieşte un model de regresie în care toate variabilele apar standardizate, cu ajutorul operatorului cunoscut, care realizează o centrare şi o normare a variabilei. De exemplu, pentru variabila Y, standardizarea se realizează astfel:

$$\widetilde{Y} = \frac{Y - \overline{Y}}{\sigma_Y} .$$

În urma standardizării, discutăm despre variaţia în unităţi de abateri standard pentru fiecare variabilă. Avantajul standardizării este acela că face posibilă compararea coeficienţilor de regresie din model, pentru că valorile standardizate ale factorilor sunt comparabile. Pentru comparaţie, se consideră valoarea estimată a coeficienţilor în modul sau în valoare absolută. Fiecare coeficient arată impactul parţial al variaţiei cu o unitate a variabilei independente standardizate asupra variabilei dependente standardizate. Valoarea coeficienţilor de regresie din modelul standardizat se interpretează ca unităţi de abateri standard pentru variabila dependentă. Cel mai mare coeficient în valoare absolută indică cea mai mare influenţă asupra variabilei dependente, iar semnul coeficientului arată sensul acestei influenţe.

BCU IAŞI CENTRALA UNIVERSITARĂ LIBRARY

Exemplul 10

Considerăm variabilele din exemplul anterior şi modelul de regresie liniară multiplă (obiectul numit *model3*) care explică variaţia natalităţii la nivelul judeţelor.

Rezultatele modelării sunt prezentate în tabelul de mai jos.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11,44504	2,02102	5,663	1,8e-06 ***
Cas	0,39949	0,19190	2,082	0,044341 *
Div	-1,19649	0,40330	-2,967	0,005248 **
Mort	-0,37392	0,10062	-3,716	0,000666 ***
Inf	0,11838	0,07231	1,637	0,110096

Coeficienţii de regresie pentru modelul în care variabilele sunt standardizate se obţin cu ajutorul funcţie *lm.beta*. Comanda utilizată şi rezultatele se prezintă astfel:

```
> lm.beta(model3)
```

Cas	Div	Mort	Inf
0,3335948	-0,3600912	-0,5679917	0,2439515

Valorile obţinute mai sus ne permit o ierarhizare a factorilor de influenţă asupra variabilei dependente astfel: cel mai important factor este mortalitatea, cu un coeficient egal cu -0,56. Urmează divorţialitatea, nupţialitatea şi mortalitatea infantilă. Acest din urmă factor are cel mai mic impact asupra natalităţii (cu un coeficient egal cu 0,24).

Interpretare

Pentru variabila nupţialitate, coeficientul de regresie din modelul cu variabile standardizate are valoarea 0,33 şi se interpretează astfel: la o creştere a nivelului nupţialităţii cu o abatere standard, nivelul mediu al natalităţii creşte cu 0,33 abateri standard, în condiţiile în care influenţa celorlalţi factori rămâne constantă.

4.6. Studiu de caz

Pentru a exemplifica demersul modelării econometrice a unui fenomen cu ajutorul modelului liniar multiplu, utilizăm datele statisticii oficiale la nivel de judeţe, pentru România, în anul 2015.

BCU IASI CENTRAL UNIVERSITY LIBRARY

Tabelul 4.1. *Repartiția județelor României după numărul pensionarilor, PIB și numărul populației din mediul rural, în 2015*

Județ	Pensionari Y	PIB (mil. lei) X_1	Populație rural X_2
Bacău	141519	12836,5	344395
Botoșani	77958	6629,9	238276
Iași	147588	20604,2	423797
Neamț	115562	8843,0	296254
Suceava	137897	12249,3	372652
Vaslui	81600	6006,5	235808
Brăila	79865	7357,1	117553
Buzău	108675	9431,1	270155
Constanța	142357	33901,5	216689
Galați	125837	12649,1	238887
Tulcea	44513	5151,6	110599
Vrancea	68842	6748,9	213476
Argeș	147326	17878,3	325560
Călărași	63257	5845,4	191860
Dâmbovița	112041	11594,5	365051
Giurgiu	57808	6027,8	196755
Ialomița	57113	6161,4	150306
Prahova	196096	32602,7	382871
Teleorman	94059	6704,6	246695
Dolj	152761	16563,3	312079
Gorj	76696	9980,6	182748
Mehedinți	58984	4748,3	138449
Olt	93406	8459,5	256822
Vâlcea	100554	8588,3	201994
Arad	101505	13820,0	191056
Caraș-Severin	75407	6831,8	132679
Hunedoara	125398	10917,1	102795
Timiș	149564	29770,5	270096
Bihor	147730	15289,6	291091
Bistrița-Năsăud	57276	6779,1	177093
Cluj	165109	29805,6	237181
Maramureș	112181	10893,6	201309
Satu Mare	76725	8106,8	186967
Sălaj	55948	5766,8	133089
Alba	84816	11009,3	140727
Brașov	142108	22014,4	156273
Covasna	44953	4723,7	109328
Harghita	73769	7104,3	177770
Mureș	134082	14445,9	274238
Sibiu	95407	13907,4	136564
Ilfov	72313	17149,5	239057

Sursa: Anuarul Statistic al României, 2016, www.insse.ro

BCU IASI

1. Prezentarea problemei

Se analizează variaţia numărului de pensionari, la nivelul judeţelor României, sub influenţa gradului de dezvoltare economică a fiecărui judeţ, măsurat prin PIB, şi a efectivului populaţiei rurale din fiecare judeţ.

2. Variabile

În exemplu, sunt utilizate următoarele variabile:

- *variabila dependentă*: numărul de pensionari (persoane), la nivel de judeţ;
- *variabile independente*: PIB (mil. de lei) şi numărul de persoane din mediul rural, pe fiecare judeţ (persoane).

Datele statistice pentru construirea modelului econometric sunt prezentate în tabelul 4.1.

3. Analiza grafică

În figura de mai jos se poate observa că fiecare variabilă independentă admite un tip de legătură liniară pozitivă cu variabila dependentă. Ne aşteptăm, aşadar, ca în modelul de regresie, coeficienţii de regresie pentru cele două variabile să fie pozitivi.

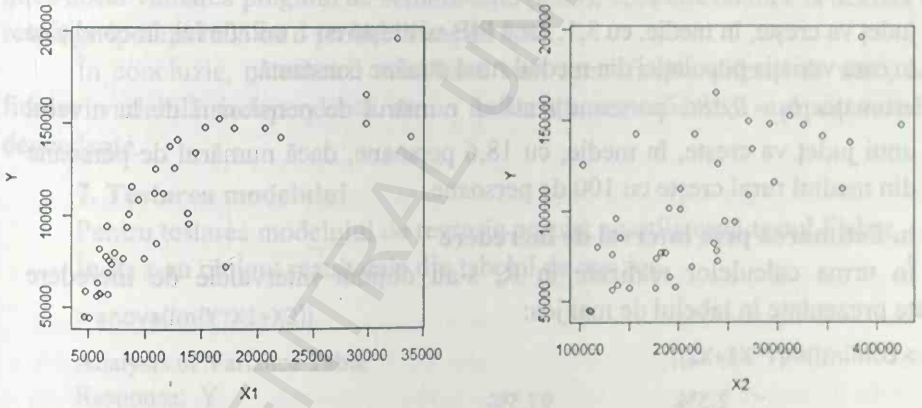


Figura 4.1. *Repartiţia judeţelor după numărul pensionarilor, PIB şi numărul persoanelor din mediul rural*

Sursa: Realizată de autor pe baza datelor din tabelul 4.1

4. Modelul de regresie

Pentru variabilele prezentate, se utilizează un model de regresie liniară multiplă cu două variabile independente. Modelul are următoarea formă:

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, unde

- Y este numărul de pensionari (persoane);
- X_1 este PIB (mil. de lei);
- X_2 este numărul de persoane din mediul rural (persoane).

BCU "CAROL I" CENTRAL UNIVERSITY LIBRARY

5. Estimarea parametrilor modelului

Prelucrarea datelor s-a realizat în R. Prin utilizarea funcției *lm* pentru cele 3 variabile, s-au obţinut următoarele rezultate:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2,240e+04	8,127e+03	2,756	0,00894 **
X1	3,103e+00	4,010e-01	7,740	2,51e-09 ***
X2	1,860e-01	3,772e-02	4,930	1,65e-05 ***

a. Estimarea punctuală a parametrilor modelului

Conform rezultatelor obţinute mai sus, modelul estimat al legăturii dintre variabilele considerate este următorul:

$$Y_X = 22400 + 3{,}1X_1 + 0{,}186X_2 .$$

Interpretare

Estimaţiile parametrilor modelului de regresie au următoarea semnificaţie:

- estimația $b_0 = 22400$ de persoane reprezintă efectivul mediu estimat al variabilei dependente, în condițiile în care variabilele independente sunt nule;
- estimația $b_1 = 3{,}1$ persoane arată că numărul de pensionari de la nivelul unui judeţ va creşte, în medie, cu 3,1, dacă PIB-ul creşte cu 1 mil. de lei, în condiţiile în care variaţia populaţiei din mediul rural rămâne constantă;
- estimația $b_2 = 0{,}186$ persoane arată că numărul de pensionari de la nivelul unui judeţ va creşte, în medie, cu 18,6 persoane, dacă numărul de persoane din mediul rural creşte cu 100 de persoane.

b. Estimarea prin interval de încredere

În urma calculelor realizate în R, s-au obţinut intervalele de încredere estimate prezentate în tabelul de mai jos:

```
> confint(lm(Y~X1+X2))
```

	2,5%	97,5%
(Intercept)	5945,2692802	3,884956e+04
X1	2,2916559	3,915098e+00
X2	0,1095946	2,623022e-01

Interpretare

Cu o probabilitate de *0,95*,

- parametrul β_0 este acoperit de intervalul: *(5945,2 ; 38849,5)*;
- parametrul β_1 este acoperit de intervalul: *(2,29 ; 3,91)*;
- parametrul β_2 este acoperit de intervalul: *(0,10 ; 0,26)*.

BCU IASI / CENTRAL UNIVERSITY LIBRARY

6. Testarea parametrilor modelului

Rezultatele testării, obținute în R, se prezintă în tabelul de mai jos:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2,240e+04	8,127e+03	2,756	0,00894 **
X1	3,103e+00	4,010e-01	7,740	2,51e-09 ***
X2	1,860e-01	3,772e-02	4,930	1,65e-05 ***

Interpretare

Pentru parametrii modelului, pe baza datelor disponibile, au fost obținute valorile calculate ale statisticii Student:

- $t_{calc} = \frac{b_0}{s_{\hat{\beta}0}} = \frac{22400}{8127} = 2,75$, pentru parametrul β_0;
- $t_{calc} = 7,74$, pentru parametrul β_1;
- $t_{calc} = 4,93$, pentru parametrul β_2.

Semnificația testului Student pentru fiecare parametru (*Sigt = 0,00*) este mai mică decât valoarea pragului de semnificație (*0,05*), ceea ce conduce la decizia de a respinge ipoteza nulă cu o probabilitate de *0,95*.

În concluzie, parametrii modelului sunt semnificativ diferiți de zero, adică fiecare variabilă independentă are o influență semnificativă asupra variabilei dependente.

7. Testarea modelului

Pentru testarea modelului de regresie propus, se utilizează testul Fisher.

În R, s-au obținut rezultatele din tabelul de mai jos.

```
> anova(lm(Y~X1+X2))
```

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	3,8080e+10	3,8080e+10	123,540	1,670e-13 ***
X2	1	7,4921e+09	7,4921e+09	24,306	1,651e-05 ***
Residuals	38	1,1713e+10	3,0824e+08		

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Interpretare

Semnificația testului Fisher are valoarea *p-value = 0,00* și este mai mică decât valoarea pragului de semnificație implicit (*0,05*), ceea ce conduce la decizia de a respinge ipoteza nulă. Rezultă că modelul liniar multiplu propus este semnificativ pentru explicarea dependenței dintre variabilele analizate.

BCU CLUJ CENTRAL UNIVERSITY LIBRARY

Observație

Testul Fisher se poate construi și pe baza estimației raportului de determinație. Acest indicator este prezentat în tabelul următor:

Multiple R-squared: 0,7955, Adjusted R-squared: 0,7848
F-statistic: 73,92 on 2 and 38 DF, p-value: 7,979e-14

Valoarea estimată a coeficientului de determinație este $R^2 = 0,79$, ceea ce înseamnă că variația variabilei dependente, numărul de pensionari, este explicată în proporție de 79% de variația simultană a celor două variabile independente, PIB și numărul de persoane din mediul rural.

8. Estimarea coeficienților de corelație

a. Coeficienții de corelație bivariată

În tabelul de mai jos, este prezentată matricea de corelație pentru variabilele incluse în model.

	Y	X1	X2
Y	1,0000000	0,8153184	0,6879001
X1	0,8153184	1,0000000	0,4469224
X2	0,6879001	0,4469224	1,0000000

Coeficienții de corelație din matricea de corelație măsoară intensitatea legăturii dintre două variabile, fără a considera interacțiunea cu celelalte variabile din model. De exemplu, între numărul mediu de pensionari și PIB, s-a estimat un coeficient de corelație bivariată de *0,81*, ceea ce indică o legătură directă și puternică.

b. Coeficienții de corelație parțială

Se poate determina un coeficient de corelație care măsoară intensitatea influenței variabilei PIB asupra variabilei dependente, număr mediu de pensionari, în condițiile în care influența variabilei independente, număr de persoane din mediu rural, este considerată constantă.

Calculul în R a condus la următorul rezultat:

	Y	X1	X2
Y	1,0000000	0,7822142	0,6245856

Coeficienții de corelație parțială dintre numărul de pensionari și cele două variabile independente sunt:

$r_{y1.2} = 0,78$; $r_{y2.1} = 0,62$.

Ambii coeficienți de corelație parțială sunt pozitivi și au același semn cu cel al coeficienților de regresie.

4.7. Teste și exerciții

Teste grilă

1. Rezultatele modelării pentru *rata natalității*, *rata mortalității* și *rata divorțialității* (rate brute, calculate la 1000 de locuitori), la nivelul județelor României, în anul 2014, sunt prezentate mai jos.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14,9077	1,1715	12,725	1,84e-15 ***
Div	-1,0703	0,4124	-2,595	0,0132 *
Mort	-0,4213	0,0817	-5,156	7,64e-06 ***

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Sunt valabile interpretările:

a) rata natalității scade, în medie, cu 1,07 unități la o creștere a ratei divorțurilor cu o unitate, în condițiile în care rata mortalității rămâne constantă;

b) rata natalității scade, în medie, cu 0,42 unități la o creștere a ratei mortalității cu o unitate;

c) variabilele independente au un impact negativ asupra celei dependente;

d) rata mortalității are o influență parțială semnificativă asupra ratei natalității, cu o probabilitate de 0,95.

2. Pentru același set de variabile de la testul anterior, se cunosc:

Multiple R-squared: 0,4261, Adjusted R-squared: 0,3967
F-statistic: 14,48 on 2 and 39 DF, p-value: 1,981e-05

Au loc enunțurile:

a) variația liniară a celor două variabile independente explică variația variabilei dependente în proporție de 39,67%;

b) coeficientul de determinație este semnificativ statistic, cu o probabilitate de 0,95;

c) modelul de regresie liniară ce explică dependența dintre variabile este semnificativ statistic, cu o probabilitate de 0,95;

d) coeficientul de determinație ajustat este mai mare decât coeficientul de determinație.

3. Rezultatele modelării pentru variabilele *salariu*, *timpul lucrat în companie* (*jobtime*, luni) și *experiența anterioară* (*prevexp*, luni), la nivelul unui eșantion de angajați, sunt prezentate în tabelul de mai jos.

BCU CLUJ CENTRAL UNIVERSITY LIBRARY

Coefficients:

	Estimate	Std. Error
(Intercept)	24332.66	6376,41
prevexp	-15,95	7,46
jobtime	143,22	77,55

Sunt valabile interpretările:

a) variabila *jobtime* are un impact semnificativ asupra variabilei dependente, cu o probabilitate de 0,90;

b) toţi parametrii modelului sunt semnificativi statistic, cu o probabilitate de 0,95;

c) nu toţi parametrii modelului sunt semnificativi statistic, cu o probabilitate de 0,95;

d) estimația cu valoarea *24332,66* arată nivelul mediu al salariului unui angajat, dacă factorii considerați în model sunt constanți.

4. Pentru datele de la testul anterior, matricea estimațiilor coeficienților de corelație bivariată se prezintă mai jos.

	salary	prevexp	jobtime
salary	1,000	-0,097	0,084
prevexp	-0,097	1,000	0,003
jobtime	0,084	0,003	1,0000

Au loc afirmațiile:

a) între *salary* și *jobtime* există o legătură liniară directă slabă;

b) între variabilele independente există o legătură liniară pozitivă de intensitate foarte slabă;

c) valoarea 1 de pe diagonală arată că variabilele analizate sunt independente;

d) coeficientul de corelație bivariată între *salary* și *prevexp* este semnificativ statistic, cu o probabilitate de 0,95.

5. Dacă dintr-un model liniar multiplu cu mai multe variabile se exclude o variabilă nesemnificativă statistic, atunci, pentru noul model:

a) coeficientul de determinație este mai mic decât în cazul modelului precedent;

b) gradele de libertate pentru componenta variației reziduale nu se modifică;

c) valorile coeficienților de regresie se modifică semnificativ.

6. În cazul unui model liniar multiplu cu 4 variabile independente:

a) numărul parametrilor din model este egal cu 5;

BCU IASI / CENTRAL UNIVERSITY LIBRARY

b) se pot determina 4 coeficienți de corelație bivariată;
c) numărul gradelor de libertate pentru testul Student asociat unui parametru este egal cu $n = -5$;
d) se pot determina 4 coeficienți de corelație parțială.

7. Pentru un model liniar multiplu cu variabile standardizate:
a) nu există ordonată la origine;
b) coeficienții de regresie sunt comparabili;
c) coeficienții de regresie pozitivi au impact mai mare decât cei negativi;
d) ordinea de importanță a predictorilor este dată de ordinea estimațiilor parametrilor modelului în valoare absolută.

8. În studiul legăturii dintre o variabilă dependentă și 2 variabile independente este posibil să existe:
a) un coeficient de determinație egal cu 1;
b) un coeficient de determinație egal cu 0;
c) doar două variabile între care există o legătură liniară semnificativă;
d) coeficienți de regresie estimați cu valori egale.

Exerciții

1. Rezultatele modelării pentru *rata natalității*, *rata mortalității* și *rata divorțialității*, la nivelul județelor României, în anul 2014, sunt prezentate mai jos.

	Estimate	Std. Error
(Intercept)	12,52587	1,95191
Div	-1,08149	0,40581
Mort	-0,33257	0,09953
Cas	0,27026	0,17874

Se cere:
a) să se identifice variabilele semnificative pentru variabila dependentă, cu o încredere de 95%;
b) să se construiască intervalele de încredere ale parametrilor modelului;
c) să se estimeze variația parțială a ratei natalității, dacă mortalitatea ar scădea cu 0,5 unități;
d) să se estimeze cu cât ar trebui modificată rata divorțialității pentru a realiza o creștere cu 0,5 unități la nivelul natalității;
e) să se scrie ecuația estimată a modelului de regresie;
f) să se estimeze nivelul mediu al natalității într-un județ care are un nivel al mortalității de 12 unități, un nivel al nupțialității de 6 unități și un nivel al divorțialității de 2 unități.

BCU CLUJ UNIVERSITY LIBRARY

2. La nivelul județelor României, în anul 2014, considerăm variabilele prezentate în tabelul de mai jos: PIB pe locuitor (*pib*, lei), rata șomajului (*som*, %), gradul de urbanizare (*urb*, %), ponderea populației vârstnice (*vârst*, %).

Județul	pib	som	urb	vârst
Bihor	26683	3,7	49,2	15,6
Bistrita-Nas,	23870	4,5	37,6	15,5
Cluj	42579	2,8	66,1	16,4
Maramures	23080	3,5	57,3	15,2
Satu Mare	23774	4,4	45,1	14,1
Salaj	26213	5,9	39,5	18,1
Alba	32669	7,3	58,2	17,9
Brasov	39954	4,3	71,6	15,2
Covasna	22710	6,5	47,5	15,6
Harghita	22917	6	42,6	16,1
Mures	26409	5,8	49,9	17
Sibiu	34769	4,6	65,9	14,8
Bacau	21182	6,8	43,1	17
Botosani	16533	5	40,6	17,8
Iasi	26281	5	45,9	14,2
Neamt	19182	6,1	35,7	19,3
Suceava	19413	6,7	41	16,3
Vaslui	15441	11,4	39,4	17,3
Braila	23733	7,4	62,1	19,5
Buzau	21483	9,7	38,4	19,8
Constanta	49564	4	68,3	14,6

Județul	pib	som	urb	vârst
Galati	24048	9,5	54,6	16,4
Tulcea	24887	5,5	46,6	16,6
Vrancea	20146	6	36,2	18,7
Arges	29698	5,9	45,9	17,1
Calarasi	19485	8,2	36,1	18,6
Dâmbovita	22690	7,9	28,6	16,5
Giurgiu	21683	7	29,1	19,8
Ialomita	22990	8,1	44	17,9
Prahova	43470	5,2	49	18,1
Teleorman	18419	11,2	32,3	24,4
Ilfov	40447	1,5	43,7	12,7
Bucuresti	87021	2	100	15
Dolj	25521	9,4	51,9	18,3
Gorj	29972	7,5	45,2	16
Mehedinti	18404	10	46,3	18,6
Olt	20094	8,2	39	19,6
Vâlcea	23530	5,6	44,6	20,3
Arad	32365	2,8	55,3	16,9
Caras-Severin	23722	4,5	53,9	18
Hunedoara	26889	6,6	74,7	18,1
Timis	42897	1,6	61,1	14,1

Pentru datele din tabelul de mai sus, se consideră ca variabilă dependentă rata șomajului. Utilizând R, se cere:

a) construiți diagramele scatterplot pentru toate perechile formate din variabila dependentă și fiecare variabilă independentă. Comentați rezultatele.

b) pentru fiecare pereche dintre variabila dependentă și o variabilă independentă, să se construiască un model de regresie simplă și să se verifice semnificația parametrilor;

c) să se construiască un model de regresie multiplă cu toate variabilele independente și să se testeze semnificația parametrilor;

d) utilizând comanda step(model, direction="backward"), să se elimine din modelul inițial toate variabilele nesemnificative. Interpretați rezultatele obținute;

e) pentru modelul final, să se estimeze modificarea ratei șomajului pentru un județ în care se realizează o creștere a PIB pe locuitor cu 5000 de lei, știind că celelalte variabile independente se păstrează constante;

f) construiți modelul cu variabile standardizate și stabiliți ordinea de importanță a factorilor semnificativi din modelul final.

3. Rezultatele modelării pentru variabilele *salariu*, *timpul lucrat în companie* (*jobtime*, luni) și *experiența anterioară* (*prevexp*, luni), la nivelul unui eșantion de angajați, sunt prezentate în tabelul de mai jos.

Response: salary

	Df	Sum Sq
prevexp	1	13102
jobtime	1	9820
Residuals	471	13562

Se cere:

a) să se calculeze valoarea estimată a coeficientului de determinație;

b) să se testeze semnificația modelului de regresie;

c) să se calculeze coeficientul de determinație ajustat;

d) să se testeze semnificația fiecărei variabile din model.

4. Să reluăm analiza prezentată la Exemplul 8. Eliminarea variabilei *rata mortalității infantile* arată că aceasta nu este semnificativă pentru variabila dependentă. Însă eliminarea variabilei presupune o modificare (scădere) a coeficientului de determinație ce are loc prin trecerea de la primul la al doilea model. Să se calculeze această diferență de nivel și să se arate că este egală cu valoarea coeficientului de determinație semi-parțială pentru variabila *rata mortalității infantile* (variabila eliminată din modelul inițial).

BCU IAȘI/CENTRUL UNIVERSITY LIBRARY

BCU IASI/CENTRAL UNIVERSITY LIBRARY

Capitolul 5

REGRESIA NELINIARĂ

Specificarea unui anumit tip de model este o etapă foarte importantă în modelarea econometrică. Uneori, teoria propune explicit tipul de model ce poate fi utiiizat în studiile empirice, alteori este vorba despre munca şi intuiţia cercetătorului care îi permit să ajungă la cea mai bună alegere. În capitolele anterioare, am discutat despre modelele liniare, care sunt considerate cele mai simple modele. Pentru modelele liniare, viteza de variaţie a variabilei dependente în raport cu cea independentă este constantă pe întreg câmpul de variaţie a variabilei predictor. Teoria şi practica economică arată însă că modelele liniare nu sunt şi cel mai frecvent utilizate. În multe situaţii, cercetarea impune apelul la modele neliniare, caz în care variaţia variabilei dependente nu depinde doar de variaţia variabilei independente, ci şi de valoarea variabilei factoriale la care se înregistrează această variaţie. Cu alte cuvinte, viteza de variaţie a variabilei dependente nu mai este constantă, ci depinde de valorile variabilei independente, de valorile variabilei dependente sau simultan de ambele. Un asemenea exemplu a fost discutat în capitolul 3 privind repartiţia angajaţilor după nivelul de educaţie şi salariu. În acest caz, este destul de clar că legătura dintre variabile este neliniară (a se vedea figura 5.1). Din punct de vedere econometric, aceste modele prezintă elemente specifice, atât în privinţa posibilităţilor de estimare a parametrilor, cât şi de aplicare practică a modelelor.

În acest capitol, prezentăm o serie de modele de regresie neliniară care permit estimarea parametrilor prin metoda celor mai mici pătrate. În această categorie se includ modelele care pot fi liniarizate, de regulă, cu ajutorul funcţiei logaritm (modele exponenţiale, putere etc.), dar şi modele precum cele polinomiale şi reciproc.

BCU / UNIVERSITY LIBRARY

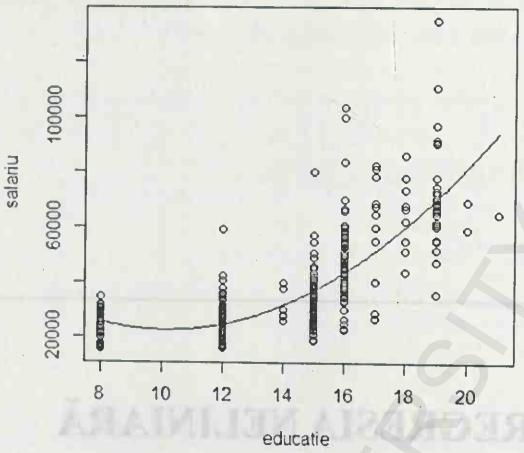


Figura 5.1. *Repartiția angajaților după salariu și nivel de educație*
Sursa: Realizată de autor (date convenționale)

5.1. Modelul log-liniar

Modelul log-liniar este un model neliniar în care variabilele apar prin funcția logaritm. Relația dintre variabilele logaritmate este de tip liniar (modelul este liniar în parametri), ceea ce permite utilizarea proprietăților modelelor liniare pentru estimarea și testarea parametrilor modelului. Acest tip de model poate fi considerat un rezultat al procesului de liniarizare, cu ajutorul funcției logaritm aplicate unui model neliniar de tip putere.

5.1.1. Estimarea modelului

Considerăm modelul de regresie cu două variabile, X, Y, de forma:

$$y_i = \beta_0 x_i^{\beta_1} e^{\varepsilon_i}.$$

Prin logaritmare, se obține modelul:

$$\ln y_i = \ln \beta_0 + \beta_1 \ln x_i + \varepsilon_i.$$

Modelul obținut este un model log-liniar, adică un model de tip liniar, în care ambele variabile apar prin funcția logaritm.

Pentru a utiliza cu ușurință proprietățile modelului liniar simplu, modelul log-liniar se poate transforma într-un model liniar, considerând notațiile:

$$y_i^* = \ln y_i;$$

$$\beta_0^* = \ln \beta_0;$$

$$\beta_1^* = \beta_1;$$

$x_i^* = \ln x_i$;

$\varepsilon_i^* = \varepsilon_i$.

Astfel, rezultă modelul: $y_i^* = \beta_0^* + \beta_1^* x_i^* + \varepsilon_i^*$.

Pentru modelul obţinut, se poate aplica metoda celor mai mici pătrate pentru estimarea parametrilor β_0^*, β_1^*. Conform rezultatelor şi proprietăţilor cunoscute pentru modelul liniar simplu, modelul nou (*) admite doi estimatori nedeplasaţi, convergenţi şi eficienţi pentru parametrii β_0^*, β_1^*.

Estimatorii au următoarele relaţii:

$$\hat{\beta}_1^* = \frac{n\sum_i \ln x_i \ln y_i - \sum_i \ln x_i \sum_i \ln y_i}{n\sum_i (\ln x_i)^2 - (\sum_i \ln x_i)^2},$$

pentru care $\hat{\beta}_1^* = \hat{\beta}_1$,

$$\hat{\beta}_0^* = \frac{1}{n}\sum_i \ln x_i - \hat{\beta}_1^* \frac{1}{n}\sum_i \ln y_i,$$

pentru care $\hat{\beta}_0^* = \ln \hat{\beta}_0$, $\hat{\beta}_0 = e^{\hat{\beta}_0^*}$.

Observaţii

1. Pentru modelul iniţial, parametrul β_1 este estimat nedeplasat cu ajutorul modelului liniar, în schimb, parametrul β_0 este estimat deplasat.

2. Pentru modelul (*), parametrul β_1 reprezintă panta dreptei sau tangenta unghiului format de dreapta de regresie cu axa OX, adică:

$$\beta_1 = \beta_1^* = \frac{dY^*}{dX^*} = \frac{d\ln Y}{d\ln X} \text{ sau}$$

$$\beta_1 = d\ln Y\big|_{d\ln X=1}.$$

Deoarece $\beta_1 = \frac{d\ln Y}{d\ln X} = \frac{\frac{dY}{Y}}{\frac{dX}{X}}$, se poate considera că parametrul reprezintă raportul dintre variaţia relativă a variabilei dependente şi variaţia relativă a variabilei independente. Această variaţie se poate exprima şi procentual:

$$\beta_1\,(\%) = d\ln Y_{(\%)}\big|_{d\ln X=1\%}.$$

BCU "SICERAL" UNIVERSITY LIBRARY

3. Viteza de variație a variabilei dependente în raport cu cea independentă nu este constantă, ci depinde de valorile celor două variabile, adică de punctul în care se calculează. Aceasta reprezintă tangenta la curba de regresie în fiecare punct:

$$\beta_1 = \frac{d\ln Y}{d\ln X} = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{dY}{dX}\frac{X}{Y}, \text{ de unde}$$

$$\frac{dY}{dX} = \beta_1 \frac{Y}{X}.$$

4. Parametrul β_0 are următoarea semnificație: este valoarea medie a variabilei dependente, când variabila independentă ia valoarea unu ($X=1$).

5.1.2. Elasticitatea

Elasticitatea unei variabile Y, în raport cu o altă variabilă X, reprezintă modificarea relativă (procentuală) a variabilei Y la o modificare relativă (procentuală) dată a lui X, de obicei mică, de o unitate.

Formalizând, elasticitatea E este dată prin relația:

$$E = \frac{\%\,\text{mod}if.Y}{\%\,\text{mod}if.X} = \frac{\frac{\Delta Y}{Y}100}{\frac{\Delta X}{X}100} = \frac{\Delta Y}{\Delta X}\cdot\frac{X}{Y},$$

unde operatorul Δ semnifică modificările sau diferențele realizate la nivelul unei variabile (operatorul diferențial).

Observații

1. Dacă modificările realizate la nivelul celor două variabile sunt mici, atunci elasticitatea se poate scrie sub forma:

$$E = \frac{dY}{dX}\frac{X}{Y} \text{ sau}$$

$$E = \frac{d\ln Y}{d\ln X}.$$

2. Pentru un model de regresie liniară simplă, elasticitatea este de forma:

$$E = \frac{dY}{dX}\frac{X}{Y} = \beta_1 \frac{X}{Y},$$

adică nu este constantă, ci depinde de raportul valorilor celor două variabile.

În practică, de obicei, se determină o elasticitate medie, pornind de la valorile medii ale celor două variabile și de la parametrul de regresie. Astfel, elasticitatea medie va fi de forma:

$$\overline{E} = \beta_1 \frac{\overline{X}}{\overline{Y}}.$$

3. Pentru modelul log-liniar, elasticitatea este tocmai parametrul β_1, adică:

$$E = \frac{d \ln Y}{d \ln X} = \beta_1.$$

Pentru acest tip de modele, elasticitatea este constantă.

Exemplu

Pentru a exemplifica demersul metodologic al modelării econometrice cu ajutorul unui model log-liniar, utilizăm un eşantion de 57 de ţări, extras de la nivel mondial, pentru care se studiază relaţia dintre variabilele *Infant mortality* (*Y*, numărul de copii decedaţi în primul an de viaţă, la 1000 de copii născuţi vii), ca variabilă dependentă, şi *Gross Domestic Product per capita* (*X*, produsul intern brut pe locuitor, exprimat în dolari), ca variabilă independentă. Datele sunt oferite de *World Bank*, pentru anul 2015.

Aşa cum arată figura 5.2 (pentru reprezentarea grafică, se utilizează funcţia *plot*), legătura dintre cele două variabile nu este liniară şi ar putea fi explicată cu ajutorul unui model log-liniar.

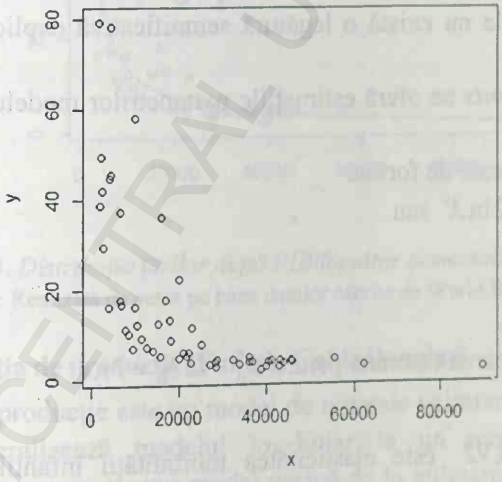


Figura 5.2. *Repartiţia bidimensională a ţărilor după PIB/locuitor şi mortalitatea infantilă, în 2015*

Sursa: Realizată de autor pe baza datelor oferite de World Bank, 2016

Pentru a estima modelul de regresie dintre cele două variabile în R, se utilizează comanda:

```
> summary(lm(log(inf)~log(gdp)))
```

BCU "ION CREANGĂ" CENTRAL UNIVERSITY LIBRARY

Se obțin rezultatele modelării:

Call:
lm(formula = log(inf) ~ log(gdp))

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11,10334	0,66392	16,72	<2e-16 ***
log(gdp)	-0,92075	0,06823	-13,49	<2e-16 ***

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 0,491 on 54 degrees of freedom
Multiple R-squared: 0,7713, Adjusted R-squared: 0,767
F-statistic: 182,1 on 1 and 54 DF, p-value: < 2.2e-16

În primul rând, observăm că estimația raportului de determinație are o valoare ridicată: $R^2 = 0,77$. Acest rezultat arată că variația variabilei dependente (rata mortalității infantile) este explicată de variația variabilei independente (PIB/locuitor) prin intermediul modelului log-liniar în proporție de 77%.

Rezultatul testului Fisher ce testează semnificația modelului sau a raportului de determinație ne conduce la decizia de a respinge ipoteza nulă (care afirmă că între cele două variabile nu există o legătură semnificativă explicată de modelul putere).

Tabelul *Coefficients* ne oferă estimaţiile parametrilor modelului şi rezultatul testării parametrilor.

Modelul estimat este de forma:

$\ln Y_X = 11,1 - 0,92 \ln X$ sau

$Y_X = e^{11,1} X^{-0,92}$.

Observație

R returnează valoarea estimată pentru $b_0^* = ln\, b_0 = 11,1$.

Interpretare

- estimaţia $b_1 = -0,92$ este elasticitatea mortalităţii infantile în raport cu produsul intern brut pe locuitor şi arată că la o creştere cu 1% a PIB/locuitor, mortalitatea infantilă scade cu 0,92%;
- estimaţia $b_0 = e^{11,1} = 66171,16$ decese la o mie de născuţi vii ne indică mortalitatea infantilă când valoarea PIB este egală cu 1$ pe locuitor. Evident, această valoare este una pur ipotetică.

Rezultă că ecuația estimată a modelului este:

BCU IASI / CENTRAL UNIVERSITY LIBRARY

$\ln Y_X = \ln 66171{,}16 - 0{,}92 \ln X$.

Testul Student bilateral indică estimații semnificative statistic pentru parametrii modelului, deoarece $Sig\,t = P(>|t|)$ are valori mai mici decât pragul de 0,05. În concluzie, se consideră că între cele două variabile există o legătură semnificativă ce poate fi modelată cu ajutorul modelului log-liniar.

Reprezentarea grafică în R a curbei estimate cu ajutorul modelului de regresie mai sus se obține cu ajutorul comenzilor:

```
> plot(inf~gdp)
> curve(66171.16*x^-0.92, add=T)
```

Se obține reprezentarea grafică de mai jos.

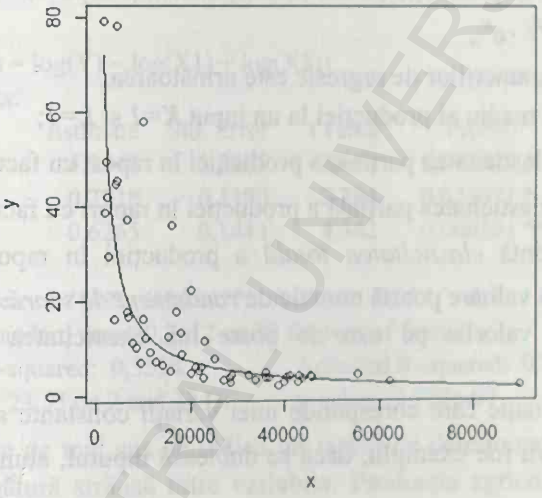


Figura 5.3. *Distribuția țărilor după PIB/locuitor și mortalitatea infantilă*
Sursa: Realizată de autor pe baza datelor oferite de World Bank, 2016

5.1.3. Funcția de producție (Funcția Cobb-Douglas)

Funcția de producție este un model de regresie neliniară multiplă de tip log-liniar, care generalizează modelul log-liniar la un număr de p variabile independente. Denumirea acestui model derivă de la utilizarea specifică a acestuia pentru a explica relația dintre producție (output) și factorii principali de producție (input).

Pentru un număr de p factori sau variabile independente, modelul are următoarea formă:

$Y = \beta_0 \cdot X_1^{\beta_1} \cdot X_2^{\beta_2} \cdot \ldots \cdot X_p^{\beta_p} \cdot e^{\varepsilon}$ sau

$$\ln Y = \ln \beta_0 + \beta_1 \ln X_1 + ... + \beta_p \ln X_p + \varepsilon .$$

În practică, se utilizează frecvent doi factori de producție: munca (*labor* – *L*) și capitalul (*K*). Modelul de regresie este următorul:

$$Y = \beta_0 \cdot X_1^{\beta_1} \cdot X_2^{\beta_2} \cdot e^{\varepsilon} \text{ sau}$$

$$Y = \beta_0 \cdot L^{\beta_1} \cdot K^{\beta_2} \cdot e^{\varepsilon} .$$

Estimarea parametrilor modelului urmează același procedeu al liniarizării, obținând un model log-liniar multiplu. Coeficienții de regresie β_j, $j = \overline{1,p}$, poartă numele de *elasticități parțiale* ale variabilei dependente în raport cu fiecare factor sau variabilă independentă.

Pentru modelul de producție cu doi factori, muncă și capital,

$$Y = \beta_0 \cdot L^{\beta_1} \cdot K^{\beta_2} \cdot e^{\varepsilon},$$

interpretarea parametrilor de regresie este următoarea:

- β_0 este nivelul mediu al producției la un input $K=1$ și $L=1$;
- β_1 reprezintă elasticitatea parțială a producției în raport cu factorul muncă;
- β_2 reprezintă elasticitatea parțială a producției în raport cu factorul capital;
- $\beta_1 + \beta_2$ reprezintă *elasticitatea totală* a producției în raport cu cei doi factori. Această valoare poartă numele de *randament de scară*.

În funcție de valorile pe care le poate lua elasticitatea totală, există următoarele situații:

- $\beta_1 + \beta_2 = 1$, situație care corespunde unei variații constante a producției în raport cu factorii (de exemplu, dacă se dublează inputul, atunci se dublează și outputul);
- $\beta_1 + \beta_2 < 1$, situație care corespunde unei variații mai reduse a producției (de exemplu, dacă se dublează inputul, atunci outputul crește într-o măsură mai mică, adică nu se realizează o dublare a producției);
- $\beta_1 + \beta_2 > 1$, situație care corespunde unei variații mai accelerate a producției, în raport cu variația factorilor.

Exemplu

Exemplificăm construirea unui model de tip funcție de producție pentru producția agricolă a României, în anul 2015, cunoscând datele statisticii oficiale pe județe, din Anuarul Statistic al României 2015.

În model, apar următoarele variabile:

- producția agricolă (mii de lei), variabilă dependentă (*Y*);

BCU "CAROL I" CENTRAL UNIVERSITY LIBRARY

- numărul mediu de salariaţi din agricultură (persoane), variabilă independentă (X_1);
- suprafaţa agricolă (ha), variabilă independentă (X_2).

Modelul funcţie de producţie este de forma:

$$Y = \beta_0 \cdot X_1^{\beta_1} \cdot X_2^{\beta_2} \cdot e^{\varepsilon}.$$

Prin liniarizare, se obţine modelul:

$$\ln Y = \ln \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \varepsilon.$$

În R, pentru variabilele menţionate, se utilizează comanda de mai jos:

```
> summary(lm(log(Y)~log(X1)+log(X2)))
```

S-au obţinut rezultatele:

```
Call:
lm(formula = log(Y) ~ log(X1) + log(X2))
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  4,1643    1,5705      2,652    0,011622 *
log(X1)      0,2718    0,1190      2,284    0,028071 *
log(X2)      0,6265    0,1443      4,342    0,000101 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 0,2612 on 38 degrees of freedom
Multiple R-squared: 0,5534,     Adjusted R-squared: 0,767
F-statistic: 23,55 on 2 and 38 DF,     p-value: 2,228e-07
```

Rezultatele de mai sus ne indică un raport de determinaţie de *0,55*, ceea ce presupune o legătură strânsă între variabile. Producţia agricolă este explicată în proporţie de 55,3% de variaţia celor doi factori luaţi în calcul, variaţie explicată cu ajutorul modelului log-liniar multiplu sau al funcţiei de producţie.

Testul Fisher ne arată că modelul funcţiei de producţie este semnificativ statistic (*p-value* < *0,05*) şi explică dependenţa dintre producţia agricolă şi factorii consideraţi: numărul mediu de salariaţi şi suprafaţa agricolă.

Tabelul *Coefficients* oferă estimaţiile parametrilor modelului şi permite scrierea ecuaţiei estimate a modelului log-liniar:

$$\ln Y_X = 4{,}16 + 0{,}27 \ln X_1 + 0{,}62 \ln X_2 \text{ sau}$$

$$\ln Y_X = \ln 64{,}07 + 0{,}27 \ln X_1 + 0{,}62 \ln X_2.$$

Testul Student asociat fiecărui parametru ne conduce la decizia de a respinge ipoteza nulă (conform căreia coeficientul de regresie nu este semnificativ statistic), cu o probabilitate de 0,95.

BCU WEST UNIVERSITY LIBRARY

Interpretare coeficienți de regresie

- estimația *4,16* ne permite determinarea estimației b_0 a parametrului β_0 prin relația: $b_0 = e^{4,16} = 64,07$. Estimația punctuală a parametrului β_0 are următoarea interpretare: valoarea producției agricole medii estimate pe județ este de 64 de mii de lei, considerând ca input o suprafață agricolă de 1 ha și 1 salariat agricol;
- estimația $b_1 = 0,27$ este elasticitatea producției agricole în raport cu factorul muncă și arată că la o creștere cu 10% a numărului de salariați din agricultură, producția agricolă crește, în medie, cu 2,7%, în condițiile în care influența factorului capital este considerată constantă;
- estimația $b_2 = 0,62$ este elasticitatea producției agricole în raport cu factorul capital și arată că la o creștere cu 10% a suprafeței agricole, producția agricolă crește, în medie, cu 6,2%, în condițiile în care influența factorului muncă este considerată constantă.

Deoarece $b_1 + b_2 < 1$, modelul de producție estimat este un model cu randament de scară descrescător.

5.2. Modele semi-logaritmice

Modelele semi-logaritmice sunt modele neliniare în care fie variabila independentă, fie variabila dependentă apare ca variabilă logaritmată. Aceste modele sunt construite, de regulă, cu scopul de a estima variația relativă sau absolută a variabilei dependente la o variație absolută sau relativă a variabilei independente.

5.2.1. Modele cu variabila dependentă logaritmată

Aceste modele sunt construite pentru studiul legăturii dintre variabile prin utilizarea modelelor matematice de tipul funcțiilor exponențiale.

Considerăm modelul de regresie de forma:

$Y = \beta_0 \beta_1^X e^{\varepsilon}$.

Prin logaritmare, se obține modelul:

$\ln Y = \ln \beta_0 + \ln \beta_1 \cdot X + \varepsilon$.

Se observă că acest model este unul liniar în care doar variabila dependentă apare logaritmată, deci este un model liniar semi-logaritmic.

Prin transformări elementare, se obține un nou model (*) de forma:

$Y^* = \beta_0^* + \beta_1^* X^* + \varepsilon^*$, unde

$Y^* = \ln Y$,

$\beta_0^* = \ln \beta_0$,

$\beta_1^* = \ln \beta_1$,

$X^* = X$,

$\varepsilon^* = \varepsilon$.

Aplicând metoda celor mai mici pătrate pentru acest nou model (*), se obţin estimatorii:

$$\hat{\beta}_1^* = \frac{n\sum_i x_i \ln y_i - \sum_i x_i \sum_i \ln y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2}, \text{ iar}$$

$\hat{\beta}_1 = e^{\hat{\beta}_1^*}$;

$\hat{\beta}_0^* = \frac{1}{n}\sum_i x_i - \hat{\beta}_1^* \frac{1}{n}\sum_i \ln y_i$, iar $\hat{\beta}_0 = e^{\hat{\beta}_0^*}$.

Modelele semi-logaritmice cu variabila dependentă transformată prin funcţia logaritm se pot utiliza în practică pentru a estima modificările relative medii ale variabilei dependente la o modificare absolută cu o unitate a variabilei independente. În funcţie de expresia matematică a modelului, se întâlnesc mai multe exemple de asemenea modele semi-logaritmice.

a. Modelul Growth

Un modelul semi-logaritmic de forma:

$\ln Y = \beta_0 + \beta_1 \cdot X + \varepsilon$,

poartă numele de model *Growth* (model de creştere) şi are la bază ecuaţia:

$Y = e^{\beta_0 + \beta_1 X + \varepsilon}$.

În acest model, parametrul β_1 reprezintă variaţia relativă a variabilei Y (variaţia în logaritm), atunci când X înregistrează o variaţie absolută cu o unitate:

$$\beta_1 = \frac{d \ln Y}{dX} = \frac{\frac{dY}{Y}}{dX} \text{ sau}$$

$\beta_1 = d \ln Y\big|_{dX=1}$.

Pentru parametrul β_1 din acest model, în interpretare, trebuie să se ţină cont de faptul că variabila dependentă înregistrează o variaţie relativă. Pentru exprimarea procentuală, este nevoie de o înmulţire a acestui coeficient cu 100:

$\beta_1 \cdot 100\,(\%) = d \ln Y_{(\%)}\big|_{dX=1}$.

Parametrul β_0 nu are o interpretare directă. Aplicându-i funcția exponențială, se obține valoarea medie a variabilei dependente, atunci când variabila independentă ia valoarea $X=0$:

$$e^{\beta_0} = Y\big|_{X=0}.$$

Pentru modelul *Growth*, elasticitatea nu este constantă și este dată prin relația:

$$E = \frac{d\ln Y}{d\ln X} = \beta_1 \cdot X.$$

Elasticitatea se poate calcula și interpreta ușor într-un anumit punct, pentru o anumită valoare a lui X.

De asemenea, viteza de variație a variabilei dependente în raport cu cea independentă nu este constantă:

$$\beta_1 = \frac{d\ln Y}{dX} = \frac{\frac{dY}{Y}}{dX}, \text{ de unde: } \frac{dY}{dX} = \beta_1 Y.$$

Dacă se consideră variația în timp a unui fenomen reprezentat de variabila Y, atunci modelul de regresie este un model de trend și are forma:

$$\ln Y = \beta_0 + \beta_1 \cdot t + \varepsilon,$$

în care t este variabila timp. Pentru acest model, elasticitatea este:

$$E = \frac{d\ln Y}{d\ln t} = \beta_1 \cdot t.$$

Parametrul β_1 oferă variația medie relativă (rata medie de variație) a variabilei Y la un moment dat.

b. Modelul Compound

O altă variantă a modelului semi-logaritmic este modelul *Compound*. Ecuația acestui model este următoarea:

$$\ln Y = \ln\beta_0 + \ln\beta_1 \cdot Y + \varepsilon.$$

Pentru acest model, are loc:

$$\ln\beta_1 = d\ln Y\big|_{dX=1}.$$

Pentru a obține variația relativă a variabilei dependente, este necesară utilizarea funcției inverse, adică a funcției exponențiale. În schimb, parametrul β_0 are o interpretare directă și oferă media lui Y pentru $X=0$.

c. Modelul Exponential

O altă variantă a modelului semi-logaritmic (numit *Exponential*) cu variabilă dependentă logaritmată este reprezentată de ecuația:

BCU IASI CENTRAL UNIVERSITY LIBRARY

$$\ln Y = \ln \beta_0 + \beta_1 \cdot X + \varepsilon .$$

Acest model este util pentru că oferă posibilitatea unei interpretări directe a fiecărui parametru. Acesta are la bază ecuaţia matematică:

$$Y = \beta_0 \cdot e^{\beta_1 X + \varepsilon} .$$

În acest caz, au loc:

$$\beta_0 = Y\big|_{X=0} ;$$

$$\beta_1 = d \ln Y\big|_{dX=1} .$$

Înmulţind valoarea lui β_1 cu 100, se obţine rata de creştere a lui Y la o modificare cu o unitate absolută a lui X:

$$\beta_1 \cdot 100(\%) = d \ln Y_{(\%)}\big|_{dX=1} .$$

Exemplu. Model Growth

Pentru a exemplifica modelul de regresie semi-logaritmic, considerăm baza de date utilizată la modelul log-liniar.

Pentru estimarea modelului semi-logaritmic, în R, utilizăm comanda:

```
> summary(lm(log(inf)~gdp))
```

Rezultatele se prezintă mai jos:

```
lm(formula = log(inf) ~ gdp
Coefficients:
             Estimate    Std. Error   t value   Pr(>|t|)
(Intercept)  3,228e+00   1,398e-01    23,090    < 2e-16 ***
gdp          -4,493e-05  4,804e-06    -9,353    6,92e-13 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 0,6343 on 54 degrees of freedom
Multiple R-squared: 0,6183,        Adjusted R-squared: 0,6112
F-statistic: 87,47 on 1 and 54 DF,    p-value: 6,918e-13
```

Rezultatele de mai sus indică o legătură relativ puternică între cele două variabile. Raportul de determinaţie este *0,61*, adică putem spune că, prin intermediul modelului, variaţia ratei mortalităţii infantile este explicată în proporţie de 61% de variaţia PIB/locuitor.

Conform testului Fisher, raportul de determinaţie este semnificativ statistic, cu o probabilitate de 0,95 (probabilitatea asociată testului, *p-value*, este mai mică de 0,05).

În figura 5.4, este prezentată repartiţia eşantionului de ţări după rata mortalităţii infantile şi nivelul PIB/locuitor. Se observă că nivelul mortalităţii infantile scade neliniar în funcţie de creşterea nivelului de dezvoltare economică.

BUCHAREST UNIVERSITY LIBRARY

Graficul se obține prin comenzile:

```
> plot(inf~gdp)
> curve(25.02*exp(-0.000049*x), add=T)
```

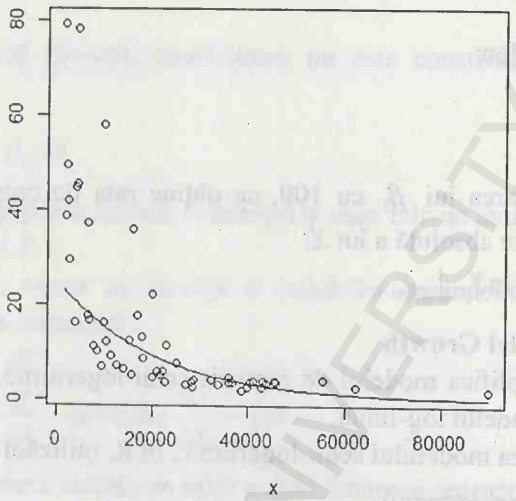


Figura 5.4. *Repartiția bidimensională a eșantionului de țări după PIB/locuitor și rata mortalității infantile*

Sursa: Realizată de autor pe baza datelor oferite de World Bank, 2016

Pe baza estimațiilor prezentate în tabelul de mai sus, se poate scrie modelul estimat:

$$\ln Y_X = 3{,}22 - 4{,}49 \cdot 10^{-5} \cdot X .$$

Interpretare

- rata mortalității infantile medie estimată, la nivelul unei țări cu PIB/locuitor nul ($X=0$), se obține din relația:
  $\ln Y = 3{,}22$.

  Rezultă $E(Y/X = 0) = e^{3,22} = 25{,}02$ decese la o mie de născuți vii.
- pentru estimația parametrului β_1, dacă se dorește exprimarea în procente a variației variabilei dependente, este valabilă interpretarea: la o creștere a nivelului PIB/locuitor cu 1000$, rata mortalității infantile scade cu 4,49% (pentru exprimarea în procente a variației variabilei dependente, s-a înmulțit valoarea coeficientului de regresie cu 100; de asemenea, pentru o variație cu 1000 de unități a variabilei independente am mai înmulțit coeficientul de regresie cu 10^3).

BCU CENTRAL UNIVERSITY LIBRARY

5.2.2. Modele cu variabila independentă logaritmată

Considerăm modelul de regresie de forma:

$Y = \beta_0 + \beta_1 \ln X + \varepsilon$.

Interesul cu privire la acest tip de model poate fi confirmat prin interpretarea parametrului de regresie β_1. Astfel, pentru acest model,

$$\beta_1 = \frac{dY}{d\ln X} = \frac{dY}{\frac{dX}{X}}, \text{ iar}$$

$\beta_1 \cdot d\ln X = dY$.

Pentru variaţii mici ale variabilelor, se poate aprecia că derivatele nu diferă semnificativ de diferenţele dintre valori şi se poate considera că panta dreptei exprimă variaţia absolută medie a variabilei dependente la o modificare relativă cu o unitate a variabilei independente.

Pentru parametrul β_1 din acest model, în interpretare, trebuie să se ţină cont de faptul că variabila dependentă înregistrează o variaţie absolută, iar variabila independentă o variaţie relativă sau o variaţie cu o unitate în logaritmi. Pentru exprimarea procentuală, este nevoie de o împărţire a acestui coeficient la 100:

$$\frac{\beta_1}{100} = dY\big|_{d\ln X = 1\%}.$$

Analog modelelor anterioare, se poate determina elasticitatea pentru un astfel de model de regresie:

$$E = \frac{d\ln Y}{d\ln X} = \beta_1 \cdot \frac{1}{Y}.$$

De asemenea, se poate observa că, pentru acest model, elasticitatea nu este constantă şi că se poate calcula uşor într-un anumit punct, pentru o anumită valoare a lui Y.

Analog, viteza de variaţie a variabilei dependente în raport cu cea independentă nu este constantă, ci depinde de valorile variabilei independente:

$$\beta_1 = \frac{dY}{d\ln X} = \frac{dY}{\frac{dX}{X}}, \text{ de unde } \frac{dY}{dX} = \beta_1 \frac{1}{X}.$$

Parametrul β_0 reprezintă valoarea medie a variabilei dependente, atunci când variabila independentă ia valoarea unu:

$\beta_0 = Y\big|_{\ln X = 0}$ sau

$\beta_0 = Y\big|_{X=1}$.

Parametrii modelului se estimează pe baza metodei celor mai mici pătrate, după relațiile cunoscute și cu respectarea condițiilor și proprietăților prezentate la modelul liniar simplu.

Exemplu

Utilizând datele din exemplul anterior, ne propunem să construim un model de regresie semi-logaritmic, în care variabila independentă apare logaritmată.

Variabilele analizate sunt:

- mortalitatea infantilă (decese la 1000 de născuți vii), ca variabilă dependentă (Y);
- PIB/locuitor ($), ca variabilă independentă (X).

În R, modelul semi-logaritmic se estimează cu ajutorul comenzii:

```
> summary(lm(inf~log(gdp)))
```

Se obțin rezultatele de mai jos.

```
Call:
lm(formula = inf ~ log(gdp)
Coefficients:
             Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)   160,47       16,06      9,995    6,97e-14 ***
log(gdp)      -14,97        1,65     -9,071    1,92e-12 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 11,87 on 54 degrees of freedom
Multiple R-squared: 0,6038,        Adjusted R-squared: 0,5964
F-statistic: 82,29 on 1 and 54 DF,    p-value: 1916e-12
```

În tabelul de rezultate, se observă că valoarea raportului de determinație este egală cu 60%, ceea ce indică o legătură puternică între variabile, iar modelul propus explică dependența dintre variabile într-o proporție ridicată. Testul Fisher arată că modelul propus pentru a explica dependența dintre mortalitatea infantilă și PIB/locuitor este semnificativ (*SigF=p-value* este mai mic decât 0,05).

Conform tabelului *Coefficients*, modelul estimat este următorul:

$Y_X = 160{,}47 - 14{,}97 \ln X$.

Interpretare

- valoarea $b_0 = 160{,}47$ decese la o mie de născuți vii este mortalitatea infantilă medie estimată pentru o țară, în condițiile în care valoarea PIB/locuitor este egală cu 1$;
- valoarea $b_1 = -14{,}97$ arată cu cât scade, în medie, mortalitatea infantilă la o creștere cu o unitate a lnPIB/locuitor. Exprimat procentual, o creștere cu 1%

a PIB/locuitor determină o scădere medie a mortalității infantile cu 0,1497 decese la 1000 de născuți vii.

Testul Student pentru fiecare parametru evidențiază că, pentru modelul considerat, parametrii sunt semnificativi statistic ($Sig\,t = Pr(>|t|)$ are valori mai mici decât 0,05).

Diagrama din figura 5.5, care prezintă legătura dintre cele două variabile, se obține în R astfel:

```
> plot(inf~gdp)
> curve(160.47-14.97*log(x), add=T)
```

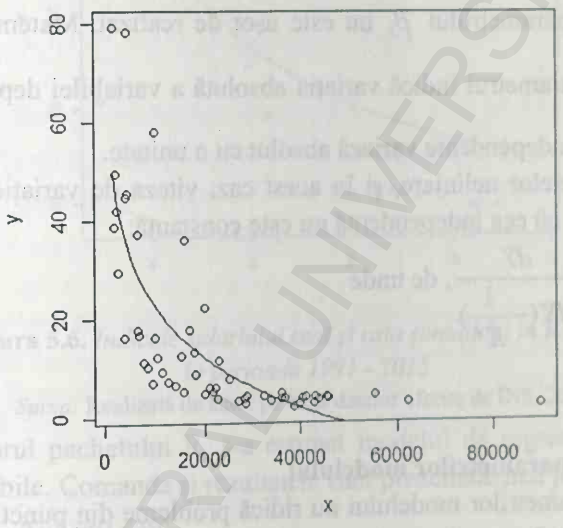


Figura 5.5. *Repartiția bidimensională a unui eșantion de țări după PIB/locuitor și mortalitatea infantilă*

Sursa: Realizată de autor pe baza datelor oferite de World Bank, 2016

5.3. Modelul reciproc. Curba lui Philips

Modelele econometrice care au la bază ecuația unei hiperbole poartă numele de *modele reciproce*. Acestea sunt modelele în care variabila independentă apare prin inversa sau prin reciproca sa.

1. Prezentarea modelului

Modelul reciproc este definit prin relația:

$$Y = \beta_0 + \beta_1 \frac{1}{X} + \varepsilon \text{ sau}$$

$$Y = \beta_0 + \beta_1 X^{-1} + \varepsilon .$$

BCU CLUJ-NAPOCA UNIVERSITY LIBRARY

Pentru acest model, parametrul β_0 reprezintă o valoare limită pe care o atinge variabila dependentă, atunci când valorile variabilei independente cresc la infinit. Valoarea lui β_0 poate fi pozitivă sau negativă.

Semnul parametrului β_1 indică sensul variației:

- dacă parametrul β_1 este pozitiv, atunci o creștere a lui X determină o descreștere a lui Y, și invers (o scădere a lui X determină o creștere a lui Y);
- dacă parametrul β_1 este negativ, atunci creșterea valorilor variabilei independente determină o creștere a valorilor variabilei dependente, și invers.

Interpretarea parametrului β_1 nu este ușor de realizat. Matematic, are loc: $\beta_1 = \frac{dY}{dX^{-1}}$, adică parametrul indică variația absolută a variabilei dependente dacă reciproca variabilei independente variază absolut cu o unitate.

Specific modelelor neliniare, și în acest caz, viteza de variație a variabilei dependente în raport cu cea independentă nu este constantă:

$$\beta_1 = \frac{dY}{dX^{-1}} = \frac{dY}{dX(-\frac{1}{X^2})}, \text{ de unde}$$

$$\frac{dY}{dX} = -\frac{1}{X^2}\beta_1.$$

2. Estimarea parametrilor modelului

Estimarea parametrilor modelului nu ridică probleme din punct de vedere al posibilității de construire a estimatorilor. Aplicarea metodei celor mai mici pătrate este posibilă, relațiile estimatorilor nefiind afectate decât prin transformarea:

$x_i \to x_i^{-1}$.

De exemplu, pentru parametrul β_1, relația este:

$$\hat{\beta}_1 = \frac{n\sum_i \frac{y_i}{x_i} - \sum_i \frac{1}{x_i}\sum_i y_i}{n\sum_i (\frac{1}{x_i})^2 - (\sum_i \frac{1}{x_i})^2}.$$

Exemplul 1. Curba lui Philips

În teoria și practica economică, a fost consacrat modelul reciproc pentru a exprima dependența dintre următoarele două variabile:

- indicele salariului real (Y), exprimat în procente (în alte modele apare rata inflației);

- rata şomajului (X), exprimată în procente.

Pentru a exemplifica modelarea econometrică cu ajutorul modelului reciproc, considerăm un set de date la nivelul României, în perioada 1991 - 2015, prezentate în figura 5.6.

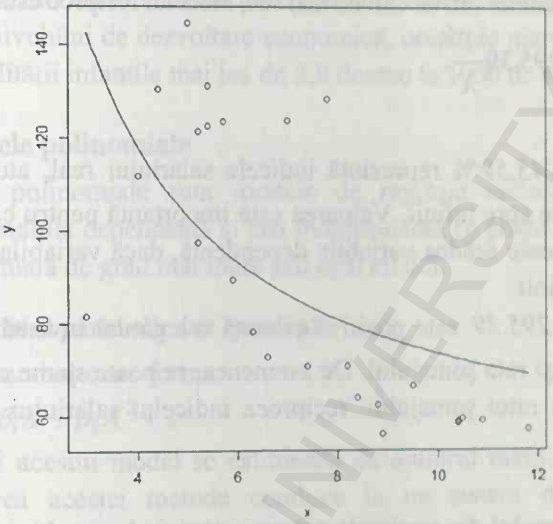


Figura 5.6. *Indicele salariului real şi rata şomajului în România, în perioada 1991 - 2015*

Sursa: Realizată de autor pe baza datelor oferite de INS, 2016

Cu ajutorul pachetului R, s-a estimat modelul de regresie reciproc pentru cele două variabile. Comanda şi rezultatele sunt prezentate mai jos.

```
> summary(lm(y~I(1/x)))

Call:
lm(formula = y ~ I(1/x))
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    43,95       14,52     3,026  0,00602 **
I(1/x)        295,39       88,11     3,353  0,00276 **
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 25,29 on 23 degrees of freedom
Multiple R-squared: 0,3283,        Adjusted R-squared: 0,2991
F-statistic: 11,24 on 1 and 23 DF,    p-value: 0,002757
```

Raportul de determinaţie sugerează că 32,8% din variaţia variabilei dependente, indicele real al salariului, este explicată de variaţia variabilei independente, rata şomajului.

BCU CLUJ CENTRAL UNIVERSITY LIBRARY

Testul Fisher conduce la decizia de a respinge ipoteza nulă (conform căreia dependența dintre variabile nu este semnificativ explicată de modelul reciproc). Cu o probabilitate de 0,95, se poate considera că modelul este semnificativ statistic (semnificația testului *p-value = 0,002*).

Conform rezultatelor din tabelul de mai sus, modelul reciproc estimat este:

$$Y_X = 43{,}95 + 295{,}39\frac{1}{X}.$$

Interpretare

- estimația $b_0 = 43{,}59\,\%$ reprezintă indicele salariului real, atunci când rata șomajului tinde spre infinit. Valoarea este importantă pentru că arată nivelul până la care poate scădea variabila dependentă, dacă variabila independentă ar crește indefinit.
- estimația $b_1 = 295{,}39$ este pozitivă și arată o legătură inversă între indicele salariului real și rata șomajului. De asemenea, se poate spune că la o creștere cu o unitate a ratei șomajului, reciproca indicelui salariului real crește cu 295,39.

Exemplul 2

Să reluăm exemplul din secțiunile anterioare privind repartiția eșantionului de țări după rata mortalității infantile și nivelul PIB/locuitor.

Așa cum arată figura 5.7, pentru cele două variabile poate fi construit un model reciproc. Rezultatele obținute în R sunt:

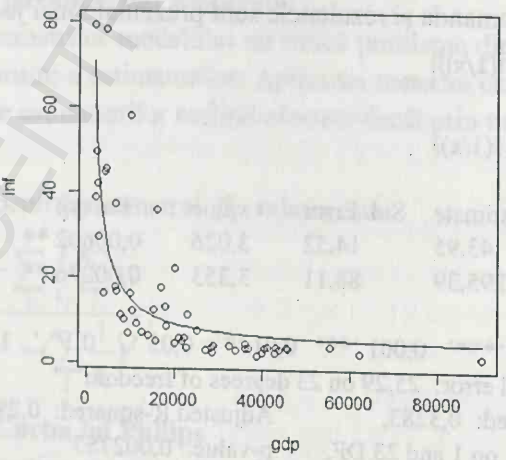


Figura 5.7. *Repartiția unui eșantion de țări după PIB/locuitor și mortalitatea infantilă*

Sursa: Realizată de autor pe baza datelor oferite de World Bank, 2016

BCU IASI CENTRUL UNIVERSITAR

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3,804e+00	2,024e+00	1,88	0,0655 .
I(1/x)	1,114e+05	1,209e+04	9,21	1,16e-12 ***

Rezultatele modelării arată că între cele două variabile legătura este inversă, iar o creştere a nivelului de dezvoltare economică, oricât de mare ar fi, nu poate să scadă rata mortalităţii infantile mai jos de 3,8 decese la 1000 de născuţi vii.

B.C.U. „M. EMINESCU" IAŞI

5.4. Modele polinomiale

Modelele polinomiale sunt modele de regresie neliniară care admit o legătură între variabila dependentă şi cea independentă ce poate fi explicată printr-o funcţie polinomială de grad mai mare sau egal cu doi.

5.4.1. Modelul parabolic sau quadratic

Modelul parabolic are forma:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon .$$

Parametrii acestui model se estimează cu ajutorul metodei celor mai mici pătrate. Aplicarea acestei metode conduce la un sistem de ecuaţii cu trei necunoscute (estimatorii parametrilor modelului), care admite trei soluţii. Sistemul de ecuaţii este de forma:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i + \hat{\beta}_2 \sum_i x_i^2 = \sum_i y_i \\ \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 + \hat{\beta}_2 \sum_i x_i^3 = \sum_i x_i y_i \\ \hat{\beta}_0 \sum_i x_i^2 + \hat{\beta}_1 \sum_i x_i^3 + \hat{\beta}_2 \sum_i x_i^4 = \sum_i x_i^2 y_i \end{cases}$$

Prin rezolvarea sistemului, se obţin relaţiile pentru cei trei estimatori, iar pe baza acestora, se obţin relaţiile de calcul pentru estimaţiile parametrilor modelului.

Ecuaţia estimată a modelului de regresie este de forma:

$$Y_X = b_0 + b_1 X + b_2 X^2 .$$

Modelul parabolic se pretează la acele aplicaţii economice care presupun o schimbare în variaţia variabilei dependente la o anumită valoare critică a variabilei independente ce corespunde unui punct de extrem (de minim sau de maxim).

Pentru identificarea punctului de extrem şi a tipului de curbă (convexă sau concavă), se utilizează condiţiile matematice specifice ecuaţiei de gradul 2:

- se obţine soluţia derivatei întâi:

$y'_{x_i} = 0$, adică $b_1 + 2b_2 x_i = 0$ sau $x_i = -\frac{b_1}{2b_2}$

- se pune condiția de extrem:

$y''_{x_i} > 0$ sau $b_2 > 0$, pentru punct de minim;

$y''_{x_i} < 0$ sau $b_2 < 0$, pentru punct de maxim.

Exemplul 1. Funcția de cost

Pentru a exemplifica un model parabolic, considerăm un set de date convenționale pentru două variabile, costul unitar (unități monetare) și producția unui bun *A* (sute bucăți), pentru un eșantion de 10 firme (figura 5.8).

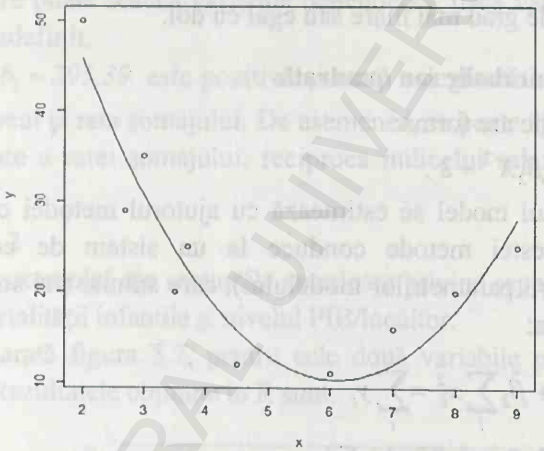


Figura 5.8. *Repartiția firmelor după costul unitar și producție*
Sursa: Realizată de autor (date convenționale)

Rezultatele de mai sus indică o legătură foarte puternică între cele două variabile, prin modelul parabolic fiind explicată variația variabilei dependente în proporție de 88,5% ($R^2 = 0,885$).

În urma aplicării testului Fisher, se ajunge la concluzia că modelul parabolic este semnificativ statistic pentru a explica dependența dintre costul unitar și producție (*p-value=0,0005*, este mai mică decât *0,05*).

Modelul estimat este de forma:

$Y_X = 89,04 - 25,795X + 2,114X^2$.

Pe baza modelului estimat, se pot face predicții și se pot stabili coordonatele punctului de minim, adică nivelul optim al producției pentru care costul unitar este minim.

Abscisa punctului de minim se obține astfel:

$$x_{opt} = -\frac{b_1}{2b_2} = \frac{25,79}{4,22} = 6,11 \text{ (sute bucăți).}$$

Această valoare corespunde unei producții de 611 bucăți din produsul A, producție la care costul unitar este minim. Valoarea estimată a costului unitar minim se obține astfel:

$$y_{\min} = 89,04 - 25,795 \cdot 6,11 + 2,114 \cdot 6,11^2 = 10,35 \text{ unități monetare.}$$

Exemplul 2. Curba Laffer

Curba Laffer este o curbă parabolică cu un punct de maxim ce prezintă legătura dintre veniturile publice și rata de taxare a veniturilor populației (numele provine de la economistul Arthur Laffer, cel care a propus acest model).

Forma curbei sugerează că la o creștere a nivelului ratei de taxare crește și venitul public; însă, după o anumită limită, veniturile încep să scadă deoarece contribuabilii vor lucra mai puțin în condițiile în care guvernul le ia tot mai mulți bani din câștigurile lor. Pentru guvern, ca să își poată maximiza veniturile, trebuie să estimeze nivelul optim de taxare. Simetria curbei parabolice sugerează că se poate utiliza un nivel de taxare mai mic pentru a atinge un nivel al veniturilor aproape de maxim.

Utilizăm un set de date convenționale pentru diferite rate de taxare a salariului (variabila *X*, procente) și nivelul veniturilor la buget (*Y*, mil. de euro), așa cum se observă din figura 5.9.

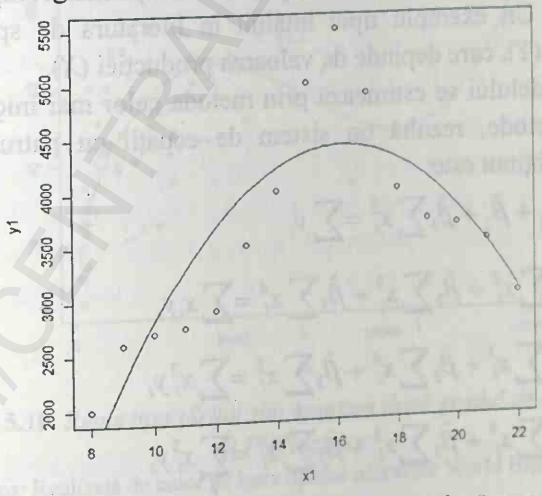


Figura 5.9. *Repartiția unui eșantion de regiuni după rata de taxare și nivelul veniturilor publice*

Sursa: Realizată de autor pe baza datelor oferite de World Bank, 2016

Rezultatele modelării în R sunt prezentate mai jos.

BCU IASI CENTRAL UNIVERSITY LIBRARY

lm(formula = $y_1 \sim x_1 + I(x_1\text{^}2)$)
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6483,566	1825,929	-3,551	0,003991 **
x	1342,673	257,029	5,224	0,000213 ***
I(x^2)	-41,351	8,498	-4,866	0,000388 ***

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 545,8 on 12 degrees of freedom
Multiple R-squared: 0,7362, Adjusted R-squared: 0,6922
F-statistic: 16,74 on 2 and 12 DF, p-value: 0,0003374

Modelul estimat pentru acest exemplu este:

$$Y_X = -6483{,}56 + 1342{,}67X - 41{,}35X^2 .$$

Nivelul de taxare care asigură maximul venitului public se estimează astfel:

$$x_{opt} = -\frac{b_1}{2b_2} = \frac{1342{,}67}{82{,}7} = 16{,}23\% .$$

5.4.2. Modelul cubic

Modelul cubic are la bază o funcţie polinomială de gradul trei şi are forma:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon .$$

Acest model este utilizat pentru a aprecia evoluţii mai complexe ale unor realităţi economice. Un exemplu tipic întâlnit în literatura de specialitate este funcţia costului total (Y), care depinde de valoarea producţiei (X).

Parametrii modelului se estimează prin metoda celor mai mici pătrate. Prin aplicarea acestei metode, rezultă un sistem de ecuaţii cu patru necunoscute. Sistemul de ecuaţii obţinut este:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i + \hat{\beta}_2 + \hat{\beta}_3 \sum_i x_i^3 = \sum_i y_i \\ \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 + \hat{\beta}_2 \sum_i x_i^3 + \hat{\beta}_3 \sum_i x_i^4 = \sum_i x_i y_i \\ \hat{\beta}_0 \sum_i x_i^2 + \hat{\beta}_1 \sum_i x_i^3 + \hat{\beta}_2 \sum_i x_i^4 + \hat{\beta}_3 \sum_i x_i^5 = \sum_i x_i^2 y_i \\ \hat{\beta}_0 \sum_i x_i^3 + \hat{\beta}_1 \sum_i x_i^4 + \hat{\beta}_2 \sum_i x_i^5 + \hat{\beta}_3 \sum_i x_i^6 = \sum_i x_i^3 y_i \end{cases}$$

Ecuaţia estimată a modelului este:

$$Y_X = b_0 + b_1 X + b_2 X^2 + b_3 X^3 .$$

BCU IASI / CENTRAL UNIVERSITY LIBRARY

Pe baza acestei ecuații, se pot determina abscisele pentru trei puncte importante ale curbei ce explică dependența dintre variabile:

- abscisele celor două puncte de extrem, ca soluții ale ecuației:
  $y'_{x_i} = 0$, adică $b_1 + 2b_2x_i + 3b_3x_i^2 = 0$;
- abscisa punctului de inflexiune, ca soluție a ecuației:
  $y'_{x_i} = 0$, adică $2b_2 + 6b_3x_i = 0$.

Exemplu

Pentru un eșantion de țări la nivel mondial, se utilizează două variabile: gradul de urbanizare (*Y*, procentul populației urbane dintr-o țară), ca variabilă dependentă, și PIB/locuitor (*X*, $), ca variabilă independentă.

Conform reprezentării grafice din figura 5.10, se observă că dependența dintre cele două variabile poate fi explicată cu ajutorul unui model cubic. Odată cu creșterea gradului de dezvoltare economică, crește și ponderea populației urbane a acelei țări. Continuarea creșterii economice poate determina și un ușor fenomen de scădere a gradului de urbanizare, prin fenomenul de migrație spre zonele rurale din preajma marilor aglomerații urbane. Creșterea economică poate antrena urbanizarea prin cooptarea acestor regiuni în zonele metropolitane.

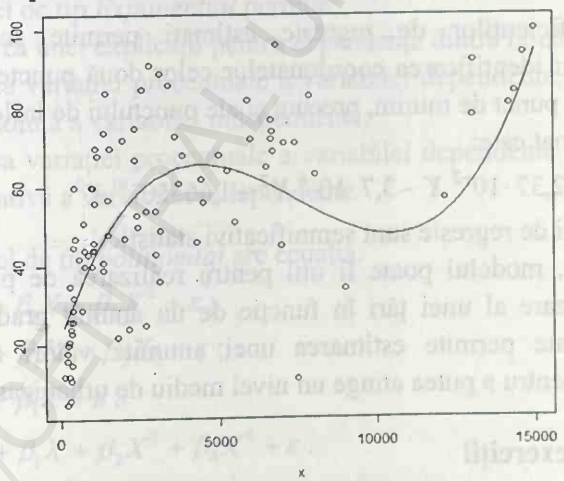


Figura 5.10. *Repartiția țărilor din eșantion după gradul de urbanizare și PIB/locuitor*

Sursa: Realizată de autor pe baza datelor oferite de World Bank, 2016

În concluzie, variația gradului de urbanizare în funcție de gradul de dezvoltare economică a statelor este mai complexă și poate fi explicată prin modelul de tip cubic.

```
> summary(lm(y~x+I(x^2)+I(x^3)))
```

```
Call:
lm(formula = y ~ x + I(x^2) + I(x^3))
Coefficients:
              Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)   2,197e+01   3,908e+00    5,622     2,24e-07 ***
x             2,377e-02   3,584e-03    6,632     2,69e-09 ***
I(x^2)       -3,710e-06   7,007e-07   -5,295     8,84e-07 ***
I(x^3)        1,661e-10   3,377e-11    4,918     4,09e-06 ***
---
Signif. codes:  0 '***'  0,001 '**'  0,01 '*'  0,05 '.'  0,1 ' ' 1
Residual standard error: 16,68 on 87 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared: 0,5268,        Adjusted R-squared: 0,5105
F-statistic: 32,28 on 3 and 87 DF,   p-value: 4,039e-14
```

Raportul de determinație indică existența unei legături intense și semnificative între variabile, după legea modelului cubic ($R^2 = 0,52$).

Testarea modelului conduce la decizia de a respinge ipoteza nulă și de a considera că modelul este semnificativ statistic ($p\text{-}value = 4 \cdot 10^{-14}$) pentru un nivel de încredere de 0,95.

Tabelul coeficienților de regresie estimați permite scrierea ecuației modelului estimat și identificarea coordonatelor celor două puncte de extrem: un punct de maxim, un punct de minim, precum și ale punctului de inflexiune.

Modelul estimat este:

$$Y_X = 21,97 + 2,37 \cdot 10^{-2} X - 3,7 \cdot 10^{-6} X^2 + 1,66 \cdot 10^{-10} X^3,$$

iar coeficienții de regresie sunt semnificativi statistic.

De asemenea, modelul poate fi util pentru realizarea de predicții privind nivelului de urbanizare al unei țări în funcție de un anumit grad de dezvoltare economică sau poate permite estimarea unei anumite valori a nivelului de dezvoltare necesar pentru a putea atinge un nivel mediu de urbanizare dorit.

5.5. Teste și exerciții

Teste grilă

1. În literatura economică, se studiază legătura dintre cantitatea de bani dintr-o economie și nivelul PIB-ului. Pentru a estima elasticitatea cantității de bani în raport cu dezvoltarea economică, se poate utiliza un model de regresie:

a) liniară simplă;

b) liniară multiplă;

BCU IASI / CENTRAL UNIVERSITY LIBRARY

c) neliniară de tip putere;
d) polinomială.

2. Un model de tip *neliniar* are ecuația:

a) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$;

b) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$;

c) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$;

d) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \ln X_3 + \varepsilon$;

e) $\ln Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \varepsilon$.

3. Diferența dintre un model liniar și unul neliniar se poate exprima cel mai bine cu ajutorul:
a) reprezentării grafice a ecuațiilor modelelor comparate;
b) coeficienților de regresie;
c) vitezei de variație a variabilei dependente în raport cu cea independentă;
d) metodei de estimare a parametrilor.

4. Un model de tip *Exponential* permite:
a) construirea unei explicații pentru dependența dintre două fenomene reale;
b) estimarea variației procentuale a variabilei dependente, dacă se cunoaște variația absolută a variabilei independente;
c) estimarea variației procentuale a variabilei dependente, dacă se cunoaște variația relativă a variabilei independente.

5. Un model de tip *polinomial* are ecuația:

a) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$;

b) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$;

c) $Y = \beta_0 + \beta_1 X + \varepsilon$;

d) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_4 X^4 + \varepsilon$.

6. Rezultatele modelării legăturii dintre variabilele PIB/loc ($) și Speranța medie de viață (ani), pentru un eșantion de țări, în anul 2007, sunt prezentate în tabelul de mai jos.

Call:
lm(formula = log(y) ~ log(x))
Coefficients:

BCU IASI UNIVERSITY LIBRARY

(Intercept) log(x)
3,27 0,095

Ecuația modelului estimat este:

a) $Y_X = 3,27 + 0,095X$;

b) $Y_X = e^{3,27} \cdot X^{0,095}$;

c) $lnY_X = 3,27 + 0,095lnX$.

7. Rezultatele modelării legăturii dintre variabilele PIB/loc ($) și mortalitatea infantilă (decese la 1000 de născuți vii), pentru un eșantion de țări, sunt prezentate în tabelul de mai jos.

Call:
lm(formula = log(y) ~ log(x))
Coefficients:
(Intercept) log(x)
7,23 -0,628

Valoarea -0,628 indică:

a) scăderea mortalității infantile la o creștere cu 1$ a PIB/loc;

b) creșterea mortalității infantile la o creștere cu 1$ a PIB/loc;

c) scăderea procentuală a mortalității infantile la o creștere cu 1% a PIB/loc.

8. Rezultatele modelării legăturii dintre variabilele PIB/loc ($) și Speranța medie de viață (ani), pentru un eșantion de țări, sunt prezentate în tabelul de mai jos.

Call:
lm(formula = y ~ log(x))
Coefficients:
(Intercept) log(x)
21,67 6,154

Sunt corecte afirmațiile:

a) ecuația estimată a modelului de regresie este: $Y = 21,67 + 6,154 \ln X$;

b) valoarea $b_1 = 6,154$ divizată cu 100 arată cu cât crește, în medie, speranța de viață la o creștere cu 1% a PIB/loc;

c) valoarea $b_0 = 21,67$ este speranța medie de viață estimată pentru o țară, în condițiile în care valoarea PIB/loc este de 0$.

9. Rezultatele modelării legăturii dintre variabilele PIB/loc ($) și numărul de calorii consumate zilnic, pentru un eșantion de țări, sunt prezentate în tabelul de mai jos.

Call:
lm(formula = log(y) ~ x)
Coefficients:
(Intercept) log(x)
7,77 $2{,}14 \cdot 10^{-5}$

Sunt corecte afirmaţiile:

a) ecuaţia estimată a modelului de regresie este: $\ln Y_X = 7{,}77 + 2{,}14 \cdot 10^{-5} \cdot X$;

b) la o creştere a PIB/loc cu 1000$, numărul de calorii consumate zilnic creşte, în medie, cu 2,14;

c) legătura dintre cele două variabile este una inversă.

Exerciţii

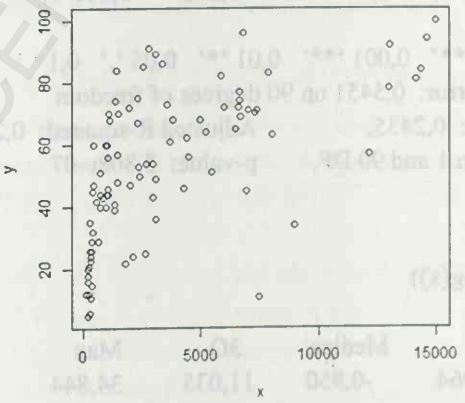
1. Prezentaţi câteva exemple de variabile economice din literatura de specialitate pentru care relaţiile de cauzalitate sunt explicate cu ajutorul unor modele neliniare.

2. Pentru două variabile economice, se dispune de un set de date reale. Pe baza acestora, s-au construit mai multe modele posibile, liniare şi neliniare. Prezentaţi câteva modalităţi de a alege cel mai bun model care să explice legătura dintre variabile, având la dispoziţie datele disponibile.

3. Care sunt asemănările şi deosebirile dintre modelele:

$$Y = \beta_0 + \beta_1 \frac{1}{X} + \varepsilon\,; \quad \frac{1}{Y} = \beta_0 + \beta_1 X + \varepsilon\,.$$

4. Pentru două variabile economice, s-a obţinut reprezentarea grafică de mai jos. Prezentaţi câteva modele posibile care ar putea explica relaţia dintre variabile.



BCU IASI/CENTRAL UNIVERSITY LIBRARY

5. În R, s-au obținut următoarele rezultate pentru două variabile *X*, *Y*. Se cere să se analizeze fiecare set de rezultate și să se propună cel mai bun model dintre cele prezentate.

Model 1
Call:
lm(formula = log(y) ~ log(x))
Residuals:

Min	1Q	Median	3Q	Max
-1,89444	-0,10997	0,04103	0,28246	0,76298

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1,24859	0,28731	4,346	3,64e-05***
log(x)	0,34129	0,03768	9,058	2,60e-14***

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 0,4532 on 90 degrees of freedom
Multiple R-squared: 0,4769, Adjusted R-squared: 0,4711
F-statistic: 82,06 on 1 and 90 DF, p-value: 2,601e-14

Model 2
Call:
lm(formula = log(y) ~ x)
Residuals:

Min	1Q	Median	3Q	Max
-1,9366	-0,2334	0,1595	0,2847	0,7863

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3,529e+00	7,786e-02	45,327	< 2e-16***
x	8,067e-05	1,499e-05	5,382	5,81e-07***

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 0,5451 on 90 degrees of freedom
Multiple R-squared: 0,2435, Adjusted R-squared: 0,2351
F-statistic: 28,96 on 1 and 90 DF, p-value: 5,808e-07

Model 3
Call:
lm(formula = y ~ log(x))
Residuals:

Min	1Q	Median	3Q	Max
-60,715	-6,964	-0,950	11,035	34,844

Coefficients:

	Estimate	Std. Error
(Intercept)	-49,957	10,464
log(x)	13,643	1,372

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 16,51 on 90 degrees of freedom
Multiple R-squared: 0,5234, Adjusted R-squared: 0,5181
F-statistic: 98,85 on 1 and 90 DF, p-value: 3,76e-16

6. Pentru modelul 1 de la exercițiul 5 se cere:
a) să se scrie ecuația estimată a modelului de regresie;
b) să se interpreteze estimațiile parametrilor;
c) să se testeze semnificația parametrilor;
d) pentru o variație a variabilei X cu 1,5%, să se estimeze variația variabilei dependente corespunzătoare.

7. Utilizând rezultatele pentru modelul 2 de la exercițiul 5, se cere:
a) să se scrie ecuația estimată a modelului de regresie;
b) să se calculeze intervale de încredere pentru estimațiile parametrilor, cu o încredere de 95%;
c) să se estimeze cu câte unități absolute ar trebui să varieze X, pentru a avea o variație relativă a variabilei dependente de 2%.

8. În cazul modelului 3 de la exercițiul 5, se cere:
a) să se scrie ecuația estimată a modelului de regresie;
b) să se testeze semnificația parametrilor, cu o încredere de 95%;
c) să se calculeze viteza de variație a variabilei dependente în raport cu cea independentă, în condițiile în care $X = 2000$.

BCU IASI CENTRALA UNIVERSITY LIBRARY

BCU IASI/CENTRAL UNIVERSITY LIBRARY

Capitolul 6

MODELE DE REGRESIE CU VARIABILE ALTERNATIVE

Modelarea econometrică se realizează cu ajutorul variabilelor numerice. Pentru a lărgi această restricţie spre variabilele nenumerice sau atributive, se construiesc variabile alternative specifice, numite şi *variabile dummy*. Aceste variabile admit doar două valori, de obicei, zero şi unu, pe baza unei convenţii sau proprietăţi. Pentru o variabilă nominală, se pot construi una sau mai multe variabile alternative, în funcţie de interesul sau scopul modelării.

În funcţie de rolul şi locul pe care îl ocupă în modelare variabilele alternative, există două clase mari de modele econometrice: modele cu variabile *dummy* independente şi modele cu variabile *dummy* dependente.

Obiectivul acestui capitol este prezentarea metodologiei modelării econometrice în condiţiile în care se utilizează variabile calitative în ambele situaţii, ca variabile independente şi ca variabile dependente.

6.1. Modele de regresie cu variabile alternative independente

În funcţie de numărul şi rolul variabilelor care apar în modelul de regresie, se pot construi două clase de modele cu variabile independente:

- modele ANOVA, care au ca variabile independente doar variabile alternative;
- modele ANCOVA, în care, ca variabile independente, se regăsesc atât variabile alternative, cât şi variabile numerice.

În secţiunile care urmează, vom nota cu:

- D, variabilele alternative sau *dummy*;
- X, variabilele independente numerice;

BCU CLUJ UNIVERSITY LIBRARY

- α_j, parametrii asociaţi variabilelor independente alternative;
- β_j, parametrii asociaţi variabilelor independente numerice.

6.1.1. Modele ANOVA

În această secţiune, prezentăm câteva tipuri de modele ANOVA utilizate mai frecvent în practică. Modelele ANOVA depind de numărul şi tipul variabilelor calitative utilizate: modele cu una sau mai multe variabile alternative, cu una sau mai multe variabile calitative nealternative.

a. Modele ANOVA cu o variabilă alternativă

Modelele ANOVA presupun transformarea variabilelor calitative în variabile alternative speciale numite variabile *dummy*, notate cu *D*. Valorile acestei variabile sunt:

- $D=1$, dacă se îndeplineşte o anumită condiţie sau proprietate pentru unităţile populaţiei;
- $D=0$, dacă nu se îndeplineşte proprietatea cerută.

Pentru cazul $D=0$, grupa de unităţi din populaţie se numeşte *grupă de referinţă*.

În modelul clasic de regresie liniară, dacă variabila X este înlocuită cu o variabilă alternativă, obţinem un model ANOVA cu o singură variabilă care este definit prin relaţia:

$$Y=\alpha_0+\alpha_1\cdot D+\varepsilon.$$

Ca medie condiţionată, regresia are următoarea expresie:

$$E(Y/D)=\begin{cases}\alpha_0, & D=0\\ \alpha_0+\alpha_1, & D=1\end{cases}$$

Interpretarea parametrilor modelului este uşor de realizat (aşa cum se observă şi din figura 6.1):

- α_0 reprezintă valoarea medie a variabilei dependente pentru acea categorie de unităţi din populaţie care nu îndeplinesc proprietatea prin care se defineşte variabila *dummy*;
- $\alpha_0+\alpha_1$ reprezintă valoarea medie a variabilei dependente pentru acea categorie de unităţi din populaţie care îndeplinesc proprietatea cerută;
- α_1 reprezintă diferenţa dintre mediile celor două categorii de unităţi din populaţie delimitate de variabila alternativă. Mai precis, este diferenţa dintre media grupei care îndeplineşte proprietatea şi media grupei care nu îndeplineşte proprietatea.

BCU BIBLIOTECA CENTRALĂ UNIVERSITARĂ

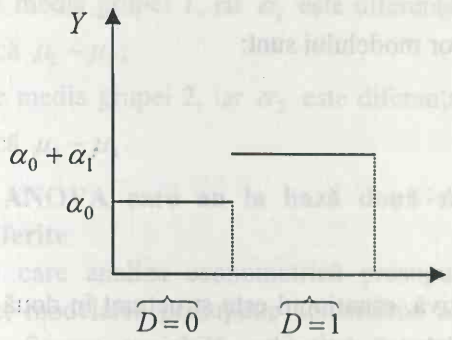


Figura 6.1. *Regresia în cazul modelului ANOVA*

Există o legătură între procedeul de analiză a varianţei (numit ANOVA) şi modelul de regresie ANOVA: ambele metode permit testarea influenţei unui factor care acţionează la două sau mai multe niveluri asupra unei variabile rezultative. Ambele procedee se rezumă la testarea egalităţii mediilor a două sau a mai multor grupe de unităţi din populaţie pentru o variabilă de interes. Avantajul abordării econometrice este acela că permite estimarea şi testarea diferenţelor mediilor grupelor construite cu ajutorul variabilelor *dummy*. Coeficienţii de regresie asociaţi acestor variabile reprezintă asemenea diferenţe, iar testele Student asociate verifică dacă diferenţele sunt semnificative statistic.

Dacă populaţia este împărţită în două grupe, se utilizează modelul de regresie:

$Y = \alpha_0 + \alpha_1 \cdot D + \varepsilon$.

Pentru acest model, notăm cu:

- μ, media populaţiei pentru variabila de interes;
- μ_1, media variabilei dependente pentru prima grupă, adică pentru $D = 0$;
- μ_2, media variabilei dependente pentru a doua grupă, adică pentru $D = 1$.

$$\mu = \frac{1}{2}(\mu_1 + \mu_2).$$

În aceste condiţii, regresia este:

$$E(Y/D) = \begin{cases} \alpha_0 = \mu_1, & D = 0 \\ \alpha_0 + \alpha_1 = \mu_2, & D = 1 \end{cases}$$

Pentru parametrii modelului, se construiesc estimatorii:

$\hat{\alpha}_0 = \hat{\mu}_1$;

$\hat{\alpha}_0 + \hat{\alpha}_1 = \hat{\mu}_2$;

BCU "SUCEAVA" TRADITIONAL UNIVERSITY LIBRARY

$\hat{\alpha}_1 = \hat{\mu}_2 - \hat{\mu}_1$.

Estimaţiile parametrilor modelului sunt:

$$a_0 = \bar{y}_1 = \frac{1}{n_1}\sum_i y_i ;$$

$$a_0 + a_1 = \bar{y}_2 = \frac{1}{n_2}\sum_i y_i ;$$

$$a_1 = \bar{y}_2 - \bar{y}_1 .$$

Prin variabila alternativă, eşantionul este structurat în două grupe de volum n_1, respectiv n_2, cu proprietatea $n_1 + n_2 = n$.

Estimarea parametrului α_1 echivalează cu estimarea diferenţei $\mu_2 - \mu_1$, iar testarea parametrului înseamnă testarea ipotezei:

$H_0: \mu_1 = \mu_2$.

b. Modele ANOVA cu două sau mai multe variabile alternative

Dacă populaţia este împărţită în mai multe grupe, cu ajutorul unei variabile nominale, utilizarea modelului ANOVA presupune construirea mai multor variabile alternative. În general, pentru o variabilă nominală cu p variante, se construiesc $p-1$ variabile alternative. Ca exemplu, prezentăm cazul unei populaţii structurate pe trei grupe, ceea ce presupune construirea a două variabile *dummy*, conform tabelului de mai jos.

Grupa	D_1	D_2
1	1	0
2	0	1
3	0	0

Pentru verificarea diferenţelor dintre cele trei grupe, se utilizează modelul ANOVA:

$Y = \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + \varepsilon$.

Pentru acest model, mediile condiţionate sunt:

$$E(Y/D) = \begin{cases} \alpha_0, & D_1 = 0,\ D_2 = 0 \\ \alpha_0 + \alpha_1, & D_1 = 1,\ D_2 = 0 \\ \alpha_0 + \alpha_2, & D_1 = 0,\ D_2 = 1 \end{cases}$$

Interpretare

- parametrul α_0 este media grupei 3, adică μ_3. Grupa 3 este grupa de referinţă, pentru că ambele variabile alternative iau valoarea 0;

- $\alpha_0 + \alpha_1$ este media grupei 1, iar α_1 este diferența dintre media grupei 1 și grupa 3, adică $\mu_1 - \mu_3$;
- $\alpha_0 + \alpha_2$ este media grupei 2, iar α_2 este diferența dintre media grupei 2 și grupa 3, adică $\mu_2 - \mu_3$.

c. Modele ANOVA care au la bază două sau mai multe variabile calitative diferite

În cazul în care analiza econometrică presupune două sau mai multe variabile calitative, modelarea presupune construirea a câte un set de variabile alternative pentru fiecare variabilă calitativă în parte, conform proprietății menționate în secțiunea precedentă.

Considerăm cazul a două variabile nominale, fiecare având două valori. Pentru fiecare variabilă este necesar să se construiască câte o variabilă alternativă diferită, în funcție de o proprietate sau condiție bine stabilită:

Condiție	D_1
A	1
A^c	0

Condiție	D_2
B	1
B^c	0

Prin A și B s-au desemnat cele două condiții care definesc variabile *dummy*, iar A^c și B^c sunt complementarele lor. Modelul de regresie ANOVA este:

$$Y = \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + \varepsilon .$$

Prin combinarea celor două criterii calitative, se construiesc patru grupe la nivelul populației analizate. În final, se obțin următoarele patru regresii care sunt, de fapt, mediile celor patru grupe din populație:

$$E(Y/D) = \begin{cases} \alpha_0, & D_1 = 0,\ D_2 = 0\ \ (A^c B^c) \\ \alpha_0 + \alpha_1, & D_1 = 1,\ D_2 = 0\ \ (AB^c) \\ \alpha_0 + \alpha_2, & D_1 = 0,\ D_2 = 1\ \ (A^c B) \\ \alpha_0 + \alpha_1 + \alpha_2, & D_1 = 1,\ D_2 = 1\ \ (AB) \end{cases}$$

Interpretare

- α_0 este media grupei 1 $(A^c B^c)$;
- $\alpha_0 + \alpha_1$ este media grupei 2 (AB^c), iar α_1 este diferența dintre media grupei 2 și media grupei 1; în același timp însă, α_1 este diferența dintre media grupei 4 și media grupei 3;

- $\alpha_0+\alpha_2$ este media grupei 3 (A^cB), iar α_2 este diferența dintre media grupei 3 și media grupei 1; similar, α_2 se obține ca diferență dintre media grupei 4 și media grupei 2;
- $\alpha_0+\alpha_1+\alpha_2$ este media grupei 4 (AB).

Analizând elementele de mai sus, se poate observa că cei 2 coeficienți de regresie asociați celor două variabile alternative au o semnificație foarte precisă:

- α_1 este diferența dintre mediile celor două grupe introduse doar de variabila D_1, conform condiției A. Adică, $\alpha_1=\mu_A-\mu_{A^c}$.
- α_2 este diferența dintre mediile celor două grupe introduse de variabila D_2, conform condiției B. Rezultă că $\alpha_2=\mu_B-\mu_{B^c}$.

În modelul ANOVA, păstrăm proprietățile specifice unui model de regresie multiplă. Coeficientul asociat fiecărei variabile arată impactul asupra variabilei dependente, când celelalte variabile sunt menținute constante. În cazul variabilelor alternative, impactul asupra variabilei dependente este măsurat ca diferență medie între grupele de unități pe care le determină respectiva variabilă. Această diferență se păstrează pentru fiecare valoare ce o ia cealaltă variabilă alternativă.

d. Modele ANOVA cu interacțiune între variabile alternative

Considerăm situația prezentată în secțiunea anterioară, adică două criterii calitative diferite pentru care s-au construit două variabile alternative diferite. Un model ANOVA cu interacțiune este de forma:

$$Y=\alpha_0+\alpha_1D_1+\alpha_2D_2+\alpha_3D_1D_2+\varepsilon.$$

La nivelul populației rămân cele patru grupe determinate de cele două criterii. Parametrii asociați primelor două variabile alternative își păstrează aceeași semnificație. Rămâne să stabilim semnificația parametrului α_3.

Mediile condiționate în acest caz sunt:

$$E(Y/D)=\begin{cases}\alpha_0, & D_1=0,\ D_2=0\ \ (A^cB^c)\\ \alpha_0+\alpha_1, & D_1=1,\ D_2=0\ \ (AB^c)\\ \alpha_0+\alpha_2, & D_1=0,\ D_2=1\ \ (A^cB)\\ \alpha_0+\alpha_1+\alpha_2+\alpha_3, & D_1=1,\ D_2=1\ \ (AB)\end{cases}$$

Se poate ușor observa că α_3 este o diferență dintre diferențele mediilor celor patru grupe: $(AB-A^cB)-(AB^c-A^cB^c)$. Acest parametru măsoară impactul interacțiunii dintre cele două variabile alternative asupra variabilei dependente.

De exemplu, dacă primul criteriu ar fi genul, iar al doilea, mediul de proveniență, iar variabila dependentă, venitul, atunci coeficientul factorului de interacțiune ar reprezenta diferența dintre diferența de salariu dintre bărbați și femei din mediul urban și diferența de salariu dintre bărbați și femei din mediul rural.

Exemplul 1

Pentru a exemplifica, construim un model de regresie de tip ANOVA pe baza datelor oficiale, oferite de ONU, 2015. Ca variabilă dependentă, se consideră vârsta mediană (*Y*, ani) a populației pe țările Europei și Asiei, în anul 2015. Variabila de structurare a populației este variabila nominală *Continent* (*D*) care admite două valori: Asia și Europa.

Grafic, repartiția pentru cele două grupe de țări după vârsta mediană se prezintă astfel:

```
> boxplot(Y~D)
```

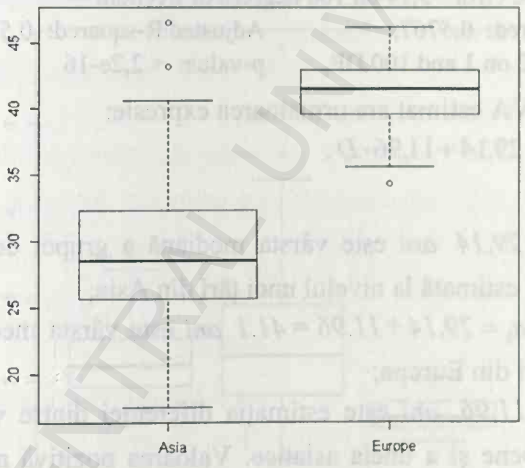


Figura 6.2. *Repartiția țărilor din Asia și Europa după vârsta mediană, 2015*
Sursa: Realizată de autor pe baza datelor oferite de ONU, 2016

Așa cum se observă din figura 6.2, ne așteptăm ca între cele două categorii de țări să existe diferențe semnificative privind vârsta mediană. Această ipoteză se verifică ușor cu ajutorul modelului ANOVA.

În R, construirea modelului presupune transformarea variabilei nominale în una alternativă. Acest lucru se realizează printr-o codare automată. Pachetul program acordă valoarea zero implicit primei valori, în ordine alfabetică, dintre valorile variabilei nominale. Așadar:

- $D = 0$, pentru o țară din Asia;

- $D = 1$, pentru o țară din Europa.

Dacă se dorește ca o anumită grupă să fie de referință, în R, se utilizează funcția *relevel*. De exemplu, putem schimba grupa de referință prin comanda:

```
> D=relevel(D, ref="Europe")
```

Rezultatele modelării se prezintă mai jos:

```
> summary(lm(Y~D))
```

```
Call:
lm(formula = Y ~ D)
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  29,1439    0,6808       42,81     <2e-16 ***
D - Europe   11,9628    1,0250       11,67     <2e-16 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 5,14 on 100 degrees of freedom
Multiple R-squared: 0,5767,     Adjusted R-squared: 0,5724
F-statistic: 136,2 on 1 and 100 DF,     p-value: < 2,2e-16
```

Modelul ANOVA estimat are următoarea expresie:

$Y = a_0 + a_1 D = 29{,}14 + 11{,}96 \cdot D$.

Interpretare

- estimația $a_0 = 29{,}14$ *ani* este vârsta mediană a grupei de referință, adică vârsta mediană estimată la nivelul unei țări din Asia;
- estimația $a_0 + a_1 = 29{,}14 + 11{,}96 = 41{,}1$ *ani* este vârsta mediană estimată la nivelul unei țări din Europa;
- estimația $a_1 = 11{,}96$ *ani* este estimația diferenței dintre vârsta mediană a unei țări europene și a uneia asiatice. Valoarea pozitivă arată că diferența este în favoarea țărilor europene. Cu alte cuvinte, în Europa, vârsta mediană este, în medie, mai mare cu 11,96 ani decât într-o țară din Asia.

Testul Student asupra parametrului α_1 ne conduce la decizia de a respinge ipoteza nulă: $H_0 : \mu_1 = \mu_2$ (vârsta mediană nu diferă semnificativ între cele două grupe de țări). Cu o probabilitate de 0,999, se poate aprecia că vârsta mediană pentru cele două grupe de țări diferă semnificativ.

Observație

Pe baza ecuației estimate $Y = a_0 + a_1 D = 29{,}14 + 11{,}96 \cdot D$, se pot scrie ecuațiile regresiilor pentru cele două categorii de țări:

- pentru țările din Asia, regresia este: $Y = 29{,}14$;

- pentru țările din Europa, regresia este: $Y = 29,14 + 11,96 = 41,1$.

Exemplul 2

Considerăm datele oferite de ONU, ca medie pentru intervalul 2010-2015, pentru variabila rată totală de fertilitate (*Y*, copii născuți de o femeie de-a lungul vieții). Ca unități, se consideră țările de la nivelul a trei continente: America Latină, Asia, Europa.

Variabila de grupare este variabila nominală *Continent*, cu cele 3 valori care determină cele 3 grupe (a se vedea figura 6.3).

În R, codarea variabilei nominale se realizează conform tabelului de mai jos. Grupa de referință va fi reprezentată de țările din America Latină.

Grupa	D_1	D_2
America Latină	0	0
Asia	1	0
Europa	0	1

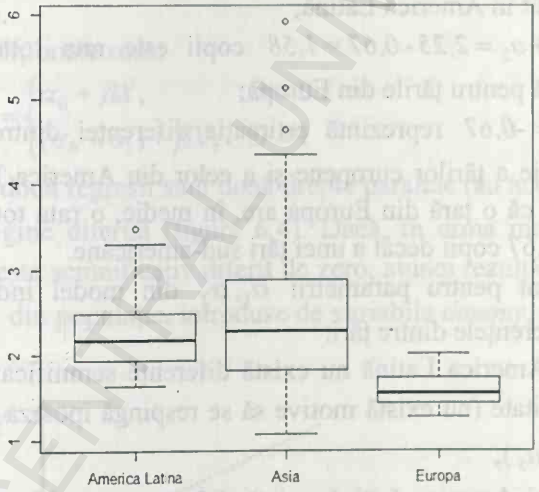


Figura 6.3. *Repartiția țărilor din America Latină, Asia și Europa după rata totală de fertilitate, 2010-2015*

Sursa: Realizată de autor pe baza datelor oferite de ONU, 2016

În R, rezultatele modelării se prezintă astfel:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2,2526	0,1037	21,719	< 2e-16 ***
D - Asia	0,2148	0,1362	1,577	0,117
D - Europa	-0,6733	0,1442	-4,669	6,95e-06 ***

BCU "ASCENTRALĂ UNIVERSITARĂ LIBRARY

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 0,6721 on 142 degrees of freedom
Multiple R-squared: 0,2456, Adjusted R-squared: 0,2349
F-statistic: 23,11 on 2 and 142 DF, p-value: 2,049e-09

Modelul estimat are următoarea ecuație:

$$Y = a_0 + a_1 D_1 + a_2 D_2 = 2{,}25 + 0{,}21 \cdot D_1 - 0{,}67 \cdot D_2 .$$

Interpretare

- estimația $a_0 = 2{,}25$ copii reprezintă rata totală de fertilitate medie estimată pentru țările din America Latină (grupa de referință);
- estimația $a_0 + a_1 = 2{,}25 + 0{,}21 = 2{,}46$ copii reprezintă rata totală de fertilitate medie estimată pentru țările din Asia;
- estimația $a_1 = 0{,}21$ copii este estimația diferenței dintre rata totală de fertilitate medie a țărilor din Asia și a celor din America Latină. Valoarea pozitivă arată că, în medie, în Asia, rata totală de fertilitate este mai mare cu 0,21 copii decât în America Latină;
- estimația $a_0 + a_2 = 2{,}25 - 0{,}67 = 1{,}58$ copii este rata totală de fertilitate medie estimată pentru țările din Europa;
- estimația $a_2 = -0{,}67$ reprezintă estimația diferenței dintre rata totală de fertilitate medie a țărilor europene și a celor din America Latină. Valoarea negativă arată că o țară din Europa are, în medie, o rata totală de fertilitate mai mică cu 0,67 copii decât a unei țări sud-americane.

Testele Student pentru parametrii α_1, α_2 din model indică următoarele concluzii privind diferențele dintre țări:

- între Asia și America Latină nu există diferențe semnificative privind rata totală de fertilitate (nu există motive să se respingă ipoteza nulă $H_0: \alpha_1 = 0$ sau $H_0: \mu_2 = \mu_3$);
- între Europa și America Latină există diferențe semnificative privind rata totală de fertilitate (se respinge ipoteza nulă $H_0: \alpha_2 = 0$ sau $H_0: \mu_1 = \mu_3$).

Observație

Pe baza ecuației estimate $Y = a_0 + a_1 D_1 + a_2 D_2 = 2{,}25 + 0{,}21 \cdot D_1 - 0{,}67 \cdot D_2$, se pot scrie ecuațiile regresiilor pentru cele trei grupe de țări:

- pentru America Latină, regresia este: $Y = 2{,}25$;
- pentru Asia, regresia este: $Y = 2{,}25 + 0{,}21 = 2{,}46$;
- pentru Europa, regresia este: $Y = 2{,}25 - 0{,}67 = 1{,}58$.

BCU "CAROL I" UNIVERSITY LIBRARY

Aceste regresii sunt drepte paralele cu axa *OX* şi oferă mediile estimate ale celor trei grupe de ţări pentru variabila dependentă.

6.1.2. Modele ANCOVA

Modelele ANCOVA conţin atât variabile *dummy*, cât şi variabile numerice. Vom considera câteva tipuri de modele: cu o variabilă alternativă şi una cantitativă, cu o variabilă cantitativă şi mai multe variabile alternative construite pe baza unei variabile nominale, cu două variabile calitative şi o variabilă numerică, modele cu interacţiune.

1. Model cu o variabilă alternativă şi o variabilă cantitativă

Modelul ANCOVA cu o variabilă alternativă şi o variabilă numerică este definit prin relaţia:

$$Y = \alpha_0 + \alpha_1 D + \beta X + \varepsilon .$$

Variabila alternativă împarte populaţia în două categorii de unităţi statistice: o grupă care îndeplineşte o proprietate ($D=1$) şi cealaltă grupă care nu respectă proprietatea ($D=0$).

Mediile condiţionate sunt:

$$E(Y/X,D) = \begin{cases} \alpha_0 + \beta X, & D = 0 \\ (\alpha_0 + \alpha_1) + \beta X, & D = 1 \end{cases}$$

Grafic, cele două regresii sunt două drepte paralele (au aceeaşi pantă β), dar cu ordonata la origine diferită (figura 6.4). Dacă, în urma modelării, rezultă că parametrul α_1 nu este semnificativ diferit de zero, atunci rezultă că între cele două categorii de unităţi din populaţie, introduse de variabila *dummy*, nu există diferenţe semnificative.

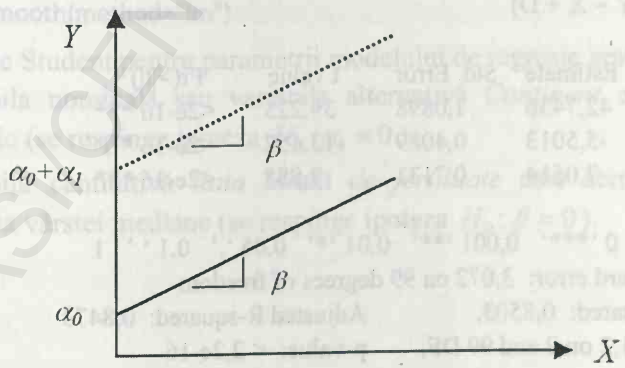


Figura 6.4. *Regresia în cazul unui model ANCOVA cu o variabilă dummy şi o variabilă cantitativă*

BCU IASI CENTRALA UNIVERSITARA

Interpretare parametri

- α_0 este nivelul mediu al variabilei dependente pentru grupa care nu respectă proprietatea impusă de variabila alternativă, în condiţiile în care $X=0$;
- $\alpha_0 + \alpha_1$ este nivelul mediu al variabilei dependente pentru grupa care respectă proprietatea impusă de variabila alternativă, în condiţiile în care $X=0$;
- α_1 este diferenţa dintre mediile celor două grupe, pentru aceeaşi valoare a lui X;
- β indică influenţa variabilei independente numerice asupra variabilei dependente. Este panta fiecărei drepte de regresie construite pentru fiecare grupă de unităţi din populaţie.

Exemplu

Considerăm datele oferite de ONU pentru ţările din Europa şi Asia, în anul 2015. Ca variabile utilizăm:

- *vârsta mediană (ani)*, variabila dependentă *(Y)*;
- *rata totală de fertilitate* (*X*, copii) şi *Continentul* (*D*), ca variabile independente. În R, variabila *Continent* este transformată automat într-o variabilă alternativă după regula:
  D=0, pentru ţările din Asia,
  D=1, pentru ţările din Europa.

Modelul ANCOVA este de forma:

$$Y = \alpha_0 + \alpha_1 D + \beta X + \varepsilon .$$

Rezultatele modelării sunt prezentate în tabelul de mai jos.

Call:
lm(formula = Y ~ X + D)
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42,7436	1,0898	39,223	<2e-16 ***
X	-5,5013	0,4089	-13,452	<2e-16 ***
D - Europa	7,0514	0,7131	9,888	<2e-16 ***

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 3,072 on 99 degrees of freedom
Multiple R-squared: 0,8503, Adjusted R-squared: 0,8473
F-statistic: 281,2 on 2 and 99 DF, p-value: < 2,2e-16

Pe baza rezultatelor din tabelul *Coefficients*, se obţine modelul estimat:

$$Y = 42{,}74 + 7{,}05 \cdot D - 5{,}5 \cdot X .$$

BCU "IOAN CENTRA" UNIVERSITY LIBRARY

Interpretare

- $a_0 = 42,74$ de ani este nivelul mediu estimat al vârstei mediane pentru o țară din Asia ($D=0$, grupa de referință), în condițiile în care nivelul fertilității este $X=0$;
- $a_0 + a_1 = 49,79$ de ani este nivelul mediu estimat al vârstei mediane pentru țările din Europa ($D=1$), în condițiile în care $X=0$;
- $a_1 = 7,05$ ani este estimația diferenței dintre vârsta mediană a țărilor din Europa și a celor din Asia, indiferent de nivelul fertilității (sau în condițiile în care variabila X este considerată constantă). Valoarea pozitivă indică o vârstă mediană mai mare pentru țările europene, în medie, cu 7,05 ani pentru același nivel al fertilității;
- $b = -5,5$ ani este scăderea nivelului mediu al vârstei mediane la nivelul unei țări, indiferent de continent (sau variabila independentă *Continentul* este considerată constantă), la o creștere a nivelului fertilității cu un copil.

Observație

Pe baza ecuației estimate $Y = 42,74 + 7,05 \cdot D - 5,5 \cdot X$, se pot scrie ecuațiile regresiilor pentru cele două categorii de persoane:

- pentru țările din Asia, regresia este:
  $Y = 42,74 - 5,5 \cdot X$;
- pentru țările din Europa, regresia este:
  $Y = 49,79 - 5,5 \cdot X$.

Cele două regresii sunt reprezentate grafic în R (vezi figura 6.5) cu ajutorul pachetului *ggplot*, folosind comanda:

```
> ggplot(an1, aes(X,Y))+ geom_point() + facet_grid(.~D)+
stat_smooth(method="lm")
```

Testele Student pentru parametrii modelului de regresie arată că:

- variabila nominală sau variabila alternativă *Continent* este semnificativă statistic (se respinge ipoteza $H_0: \alpha_1 = 0$);
- variabila cantitativă *Rata totală de fertilitate* este semnificativă pentru variația vârstei mediane (se respinge ipoteza $H_0: \beta = 0$).

BCU IAȘI - CENTRAL UNIVERSITY LIBRARY

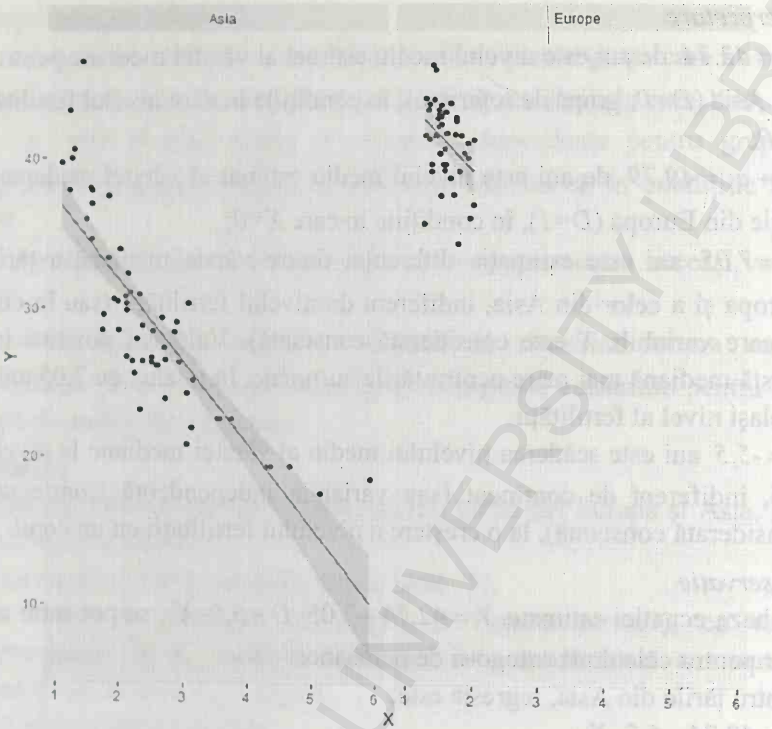


Figura 6.5. *Regresiile pentru cele două grupe de ţări*
Sursa: Realizată de autor pe baza datelor oferite de ONU, 2016

2. Model cu o variabilă cantitativă şi mai multe variabile dummy, pentru o variabilă nominală cu mai multe categorii

Considerăm, de exemplu, o variabilă nominală cu trei valori. Pentru a face distincţia între cele trei grupe de unităţi din populaţie, se construiesc două variabile alternative, conform tabelului de mai jos:

Grupa	D_1	D_2
1	1	0
2	0	1
3	0	0

Modelul de regresie ANCOVA cu o variabilă cantitativă şi mai multe variabile *dummy*, construite pe baza unei variabile nominale, are următoarea expresie:

$$Y = \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + \beta X + \varepsilon .$$

Pentru acest model, se obţin trei regresii, care au expresiile:

BCU IASI / CENTRAL UNIVERSITY LIBRARY

$$E(Y/X,D_1,D_2)=\begin{cases}\alpha_0+\beta X, & D_1=0,\ D_2=0\\ (\alpha_0+\alpha_1)+\beta X, & D_1=1,\ D_2=0\\ (\alpha_0+\alpha_2)+\beta X, & D_1=0,\ D_2=1\end{cases}$$

Parametrii modelului au următoarea semnificaţie:

- α_0 este media variabilei dependente pentru grupa de referinţă (grupa 3 de unităţi din populaţie), când $X=0$;
- α_1 este diferenţa dintre media grupei 1 şi a grupei 3, pentru variabila dependentă, pentru o anumită valoare a lui X (sau se consideră că X este constantă);
- α_2 este diferenţa dintre media grupei 2 şi a grupei 3, pentru variabila dependentă, când X este constantă;
- β este variaţia variabilei dependente la o variaţie cu o unitate a variabilei cantitative X, pentru o valoare anume a variabilei nominale.

Prin modelare, se obţin trei drepte de regresie paralele, câte una pentru fiecare dintre cele trei categorii de populaţie determinate de variabila nominală. Diferenţele dintre regresii sunt date de ordonata la origine, panta fiind aceeaşi.

Exemplu

Utilizăm baza de date de la exemplul anterior (analiza legăturii dintre vârsta mediană, rata totală de fertilitate şi continentul, în anul 2015), la care mai adăugăm o grupă de ţări din America Latină.

În R, codarea variabilei nominale se realizează astfel:

Grupa	D_1	D_2
America Latină	0	0
Asia	1	0
Europa	0	1

Pentru modelul de regresie ANCOVA, $Y=\alpha_0+\alpha_1D_1+\alpha_2D_2+\beta X+\varepsilon$, s-au obţinut estimaţiile din tabelul *Coefficients* de mai jos.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	43,92847	1,13824	38,593	< 2e-16 ***
X	-5,96655	0,44300	-13,469	< 2e-16 ***
D - Asia	-0,04272	0,72518	-0,059	0,953
D - Europa	6,60138	0,81762	8,074	2,72e-13 ***

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Residual standard error: 3,548 on 141 degrees of freedom

Multiple R-squared: 0,7815, Adjusted R-squared: 0,7769
F-statistic: 168,1 on 3 and 141 DF, p-value: < 2,2e-16

Modelul estimat are relaţia:

$Y = 43{,}92 - 0{,}04 \cdot D_1 + 6{,}6 \cdot D_2 - 5{,}96 \cdot X$.

Pe baza ecuaţiei estimate, se pot scrie ecuaţiile regresiilor pentru cele trei categorii de persoane:

- pentru ţările din America Latină, regresia este: $Y = 43{,}92 - 5{,}96 \cdot X$;
- pentru ţările din Asia, regresia este: $Y = 43{,}88 - 5{,}96 \cdot X$;
- pentru ţările din Europa, regresia este: $Y = 50{,}52 - 5{,}96 \cdot X$.

Cele 3 regresii sunt reprezentate grafic în figura 6.6.

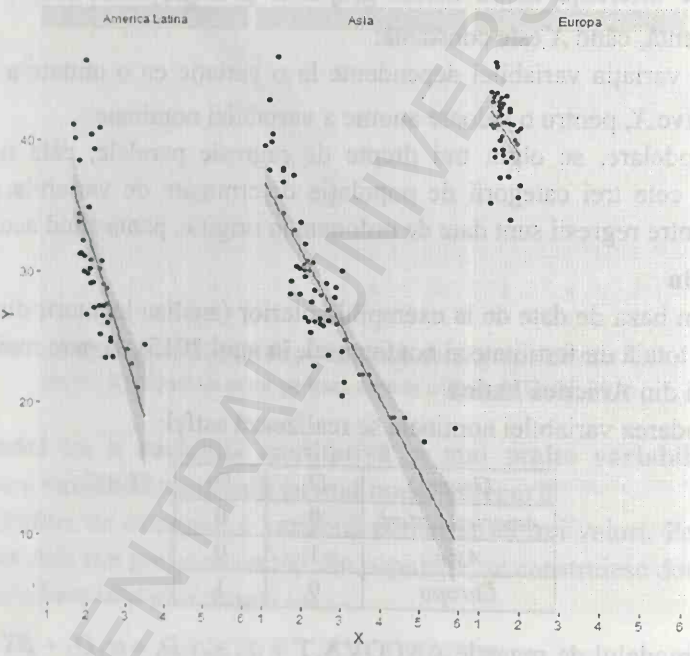


Figura 6.6. *Regresiile pentru cele trei grupe de ţări*
Sursa: Realizată de autor pe baza datelor oferite de ONU, 2016

Interpretare

- $a_0 = 43{,}92$ de ani este nivelul mediu estimat al vârstei mediane pentru ţările din America Latină, în condiţiile în care $X=0$;
- $a_1 = -0{,}04$ ani este diferenţa dintre mediile estimate pentru variabila dependentă la nivelul ţărilor din Asia şi al celor din America Latină,

BCU KRS CENTRAL UNIVERSITY LIBRARY

considerând constant nivelul fertilității. Valoarea negativă indică o diferență în defavoarea țărilor din Asia;

- $a_2 = 6,6$ este diferența dintre mediile estimate pentru variabila dependentă la nivelul țărilor din Europa și al celor din America Latină, considerând constant nivelul fertilității. Țările europene au, în medie, o vârstă mediană mai mare decât cea a țărilor Americii Latine cu 6,6 ani, pentru același nivel al fertilității;
- $b = -5,96$ ani reprezintă scăderea medie a vârstei mediane, dacă nivelul fertilității crește cu un copil, la nivelul unui continent specificat;
- $a_0 + a_1 = 43,92 - 0,04 = 43,88$ de ani este nivelul mediu estimat al vârstei mediane la nivelul țărilor din Asia, dacă $X=0$;
- $a_0 + a_2 = 43,92 + 6,6 = 50,52$ de ani este nivelul mediu estimat al vârstei mediane la nivelul țărilor din Europa, dacă $X=0$.

3. Model cu o variabilă alternativă și două variabile cantitative

Un model de acest tip este reprezentat de ecuația:

$Y = \alpha_0 + \alpha_1 D_1 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

În acest caz, pentru valorile variabilei alternative, rezultă două regresii:

$$E(Y / X_1, X_2, D) = \begin{cases} \alpha_0 + \beta_1 X_1 + \beta_2 X_2, & D = 0 \\ (\alpha_0 + \alpha_1) + \beta_1 X_1 + \beta_2 X_2, & D = 1 \end{cases}$$

Interpretare parametri

- α_0 este nivelul mediu al variabilei dependente pentru grupa de referință, dacă influența variabilelor independente numerice este considerată nulă;
- α_1 este diferența dintre media celor două grupe de unități delimitate de variabila dummy, în condițiile în care influența celor două variabile independente este constantă;
- β_1, respectiv β_2, arată influența parțială a variabilei numerice corespunzătoare asupra variabilei dependente (se consideră că cealaltă variabilă este constantă și că variația este aceeași pentru fiecare grup definit de variabila alternativă).

4. Model cu două variabile calitative și o variabilă cantitativă

Analizăm situația a două variabile calitative alternative ce structurează populația în patru grupe de unități. În acest caz, se utilizează două variabile alternative (D_1, D_2) care determină câte 2 grupe după fiecare variabilă nominală în parte.

Ecuația acestui tip de model este:

$$Y = \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + \beta X + \varepsilon .$$

Pentru valorile celor două variabile alternative, rezultă patru regresii ce corespund celor 4 grupe:

$$E(Y/X, D_1, D_2) = \begin{cases} \alpha_0 + \beta X, & D_1 = 0, \ D_2 = 0 \\ (\alpha_0 + \alpha_1) + \beta X, & D_1 = 1, \ D_2 = 0 \\ (\alpha_0 + \alpha_2) + \beta X, & D_1 = 0, \ D_2 = 1 \\ (\alpha_0 + \alpha_1 + \alpha_2) + \beta X, & D_1 = 1, \ D_2 = 1 \end{cases}$$

Exemplu

Considerăm datele oferite de ONU privind nivelul vârstei mediane şi al ratei totale de fertilitate pentru un eşantion de ţări grupate pe continente şi pe nivelul de dezvoltare economică.

În R, pentru variabila *Continent*, este codată variabila *dummy* care ia valorile $D_1=0$, pentru ţările din Asia, şi $D_1=1$, pentru ţările din Europa. Similar, pentru variabila *nivel de dezvoltare*, este codată o variabilă cu valorile $D_2=0$, pentru ţările dezvoltate, iar $D_2=1$, pentru ţările în curs de dezvoltare.

Pentru modelul ANCOVA $Y = \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + \beta X + \varepsilon$, în R, s-au obţinut rezultatele:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	43,1554	1,3172	32,764	< 2e-16 ***
X	-5,5202	0,4118	-13,407	< 2e-16 ***
D1 - Europe	6,8822	0,7766	8,862	3,56e-14 ***
D2 - Less	-0,4161	0,7418	-0,561	0,576

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Residual standard error: 3,082 on 98 degrees of freedom

Multiple R-squared: 0,8508, Adjusted R-squared: 0,8462

F-statistic: 186,3 on 3 and 98 DF, p-value: < 2,2e-16

Ecuaţia estimată a modelului este:

$$Y = 43{,}15 + 6{,}88 \cdot D_1 - 0{,}41 \cdot D_2 - 5{,}52 \cdot X .$$

Interpretare

Pe baza estimaţiilor coeficienţilor de regresie, se obţin:

- $a_0 = 43{,}15$ ani este nivelul mediu estimat al vârstei mediane a populaţiei la nivelul unei ţări din Asia, din grupa dezvoltată, în condiţiile în care $X=0$;
- $a_0 + a_1 = 50{,}03$ ani este nivelul mediu estimat al vârstei mediane a populaţiei la nivelul unei ţări din Europa, din grupa dezvoltată, pentru $X=0$;

BCU "SIMION STOILOW" UNIVERSITY LIBRARY

- $a_0 + a_2 = 42,74$ de ani este nivelul mediu estimat al vârstei mediane a populației la nivelul unei țări din Asia, din grupa în curs de dezvoltare, pentru $X=0$;
- $a_0 + a_1 + a_2 = 49,62$ de ani este nivelul mediu estimat al vârstei mediane a populației la nivelul unei țări din Europa, din grupa în curs de dezvoltare, pentru $X=0$;
- $a_1 = 6,88$ ani este nivelul mediu estimat al diferenței dintre vârsta mediană la nivelul unei țări din Europa și a uneia din Asia, considerând constante celelalte variabile independente;
- $a_2 = -0,41$ ani este nivelul mediu estimat al diferenței dintre vârsta mediană la nivelul unei țări în curs de dezvoltare și a uneia dezvoltate, considerând constante celelalte variabile independente;
- $b = -5,52$ ani este scăderea medie a vârstei mediane a populației la o creștere a nivelului fertilității cu un copil la o femeie, indiferent de grupa de țară.

Cele 4 regresii corespunzătoare celor 4 grupe de țări sunt următoarele:

- $Y = 43,15 - 5,52 \cdot X$, pentru țările dezvoltate din Asia;
- $Y = 50,03 - 5,52 \cdot X$, pentru țările dezvoltate din Europa;
- $Y = 42,74 - 5,52 \cdot X$, pentru țările în curs de dezvoltare din Asia;
- $Y = 49,62 - 5,52 \cdot X$, pentru țările în curs de dezvoltare din Europa.

Testele Student aplicate parametrilor modelului arată că între țările dezvoltate și cele slab dezvoltate nu există diferențe semnificative privind nivelul vârstei mediane, în timp ce între continente există asemenea diferențe. De asemenea, se poate aprecia că rata totală de fertilitate are un impact parțial semnificativ asupra variabilei dependente.

6.1.3. Aplicații ale modelelor ANOVA și ANCOVA

Modelele de tip ANOVA și ANCOVA pot fi utilizate pentru rezolvarea unor probleme clasice de inferență statistică. Prin specificul lor, aceste modele sunt folosite în scopul rezolvării unor probleme tipice modelării econometrice, cum ar fi: analiza unui model de regresie pe porțiuni, în funcție de o valoare critică; compararea a două sau mai multe modele de regresie; testarea homoscedasticității unui model de regresie; testarea sezonalității unei serii de timp etc.

În continuare, prezentăm câteva aplicații din cele menționate mai sus.

1. Analiza unui model de regresie pe porţiuni, în funcţie de o valoare critică

Există situaţii când, la nivelul repartiţiei după două variabile statistice, variaţia variabilei dependente ar putea suporta o modificare importantă începând cu o valoare critică x_0 a variabilei independente.

Considerăm că relaţia dintre cele două variabile este explicată de un model de regresie liniară simplă:

$$Y = \beta_0 + \beta_1 X + \varepsilon .$$

Valoarea critică x_0 ar putea conduce la următoarea situaţie:

- dacă $x_i < x_0$, modelul de regresie este de forma:

$$Y = \gamma_1 + \lambda_1 X + \varepsilon_1 ;$$

- dacă $x_i > x_0$, modelul de regresie îşi modifică parametrii:

$$Y = \gamma_2 + \lambda_2 X + \varepsilon_2 .$$

Pentru a testa dacă valoarea critică x_0 conduce într-adevăr la o schimbare importantă a variaţiei variabilei dependente, se construieşte un model ANCOVA de forma:

$$Y = \beta_0 + \beta_1 X + \beta_2 (X - x_0) D + \varepsilon , \text{ unde}$$

- $D = 0$, dacă $x_i < x_0$;
- $D = 1$, dacă $x_i \geq x_0$.

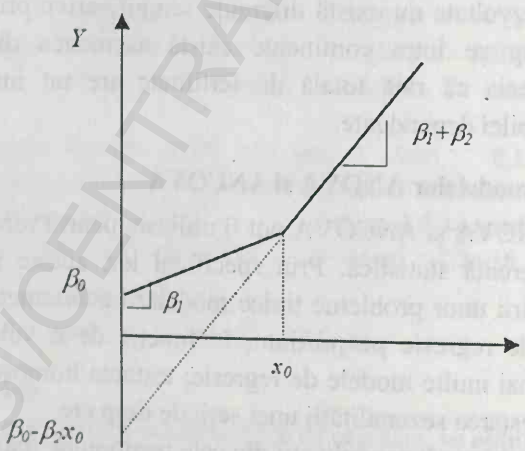


Figura 6.7. *Regresia în raport cu o valoare critică* x_0

În aceste condiţii, regresia sau media condiţionată este de forma (vezi figura 6.7):

BCU IASI CENTRAL UNIVERSITY LIBRARY

$$E(Y/X,D)=\begin{cases}\beta_0+\beta_1X, & D=0\\ (\beta_0-\beta_2x_0)+(\beta_1+\beta_2)X, & D=1\end{cases}$$

În funcție de semnificația parametrilor din ecuația de mai sus, întâlnim 2 cazuri:

a. Două regresii diferite

Dacă se resping ipotezele $H_0:\beta_0=0$, $H_0:\beta_2=0$, atunci există două regresii diferite, adică pentru valori mai mari decât valoarea critică x_0, se modifică parametrii modelului de regresie. Cele două regresii sunt:

$Y=\gamma_1+\lambda_1X+\varepsilon_1$;

$Y=\gamma_2+\lambda_2X+\varepsilon_2$,

unde:

$\gamma_1=\beta_0$, $\lambda_1=\beta_1$;

$\gamma_2=\beta_0-\beta_2x_0$, $\lambda_2=\beta_1+\beta_2$.

b. O singură regresie

Dacă se acceptă ipoteza nulă $H_0:\beta_2=0$, atunci există o singură regresie, deci valoarea critică x_0 nu produce o modificare importantă a variației pentru variabila dependentă. În acest caz, regresia este:

$Y=\beta_0+\beta_1X+\varepsilon$.

Exemplu

Considerăm un eșantion de 102 țări repartizate după vârsta mediană și rata totală de fertilitate (date oferite de ONU, 2015), conform figurii 6.8. În ipoteza că între cele 2 variabile există o legătură liniară, se poate verifica dacă viteza de variație a variabilei dependente (vârsta mediană) se modifică după pragul de $x_0=2{,}1$ copii la o femeie, limită a ratei de fertilitate numită prag de înlocuire a unei generații.

Verificarea ipotezei de modificare a vitezei de variație după pragul menționat se poate realiza cu ajutorul unui model ANCOVA în care variabila *dummy* se construiește astfel:

- $D=0$, dacă $x_i<2{,}1$;
- $D=1$, dacă $x_i\geq 2{,}1$.

Pasul următor este estimarea modelului de regresie de forma:

$Y=\beta_0+\beta_1X+\beta_2(X-x_0)D+\varepsilon$.

BCU CENTRAL UNIVERSITY LIBRARY

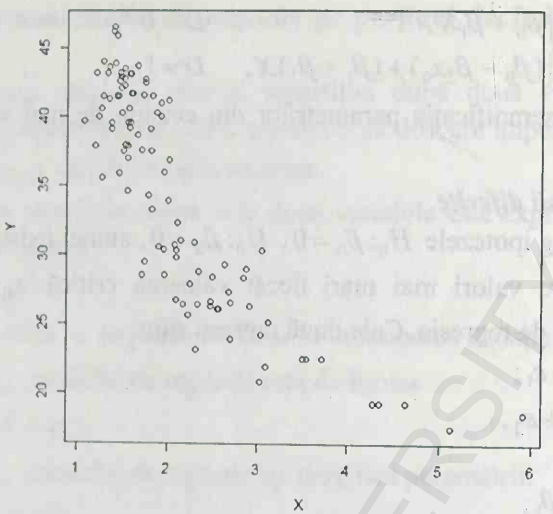


Figura 6.8. *Repartiția țărilor după vârsta mediană și rata totală de fertilitate, în 2015*

Sursa: Realizată de autor pe baza datelor oferite de ONU, 2016

Rezultatele modelării și reprezentarea grafică a celor 2 regresii sunt prezentate mai jos.

Call:
lm(formula = Y ~ X + I((X - 2.1) * D))
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	63,174	2,308	27,375	< 2e-16 ***
X	-15,399	1,333	-11,548	< 2e-16 ***
I((X - 2.1)*D)	10,619	1,717	6,185	1,4e-08 ***

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 3,678 on 99 degrees of freedom
Multiple R-squared: 0,7854, Adjusted R-squared: 0,7811
F-statistic: 181,2 on 2 and 99 DF, p-value: < 2,2e-16

Testul Student pentru parametrul β_2 conduce la decizia de a respinge ipoteza că valoarea critică ($x_0 = 2,1$ *copii*) nu produce o modificare importantă a vitezei de variație pentru variabila dependentă după această limită.

Așa cum se observă din figura 6.9, cele două regresii au pante sau viteze de variație diferite. Ecuațiile celor două regresii sunt următoarele:

$Y = 63,17 - 15,39X$, pentru $X < 2,1$;

$Y = (63,79 - 10,61 \cdot 2,1) + (-15,39 + 10,61)X = 40,87 - 4,78X$, pentru $X \geq 2,1$.

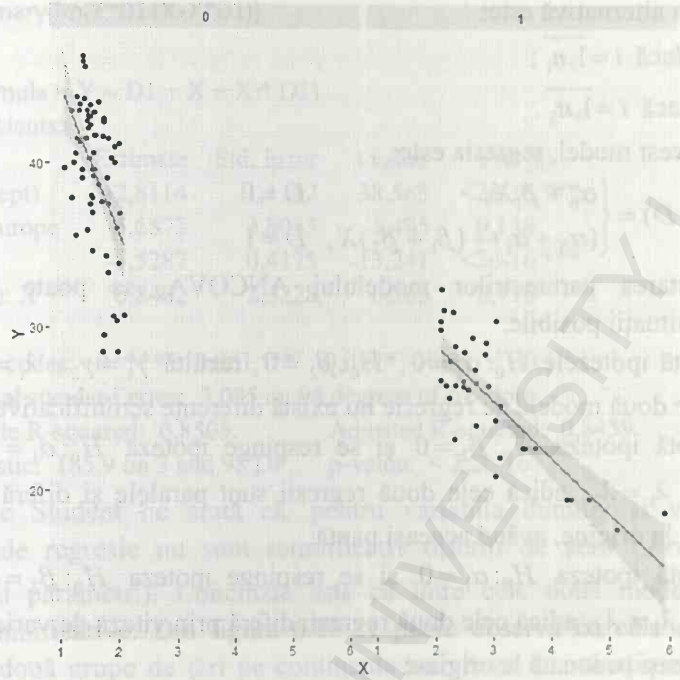


Figura 6.9. *Regresiile pentru cele două grupe de țări în funcție de valoarea critică $x_0 = 2{,}1$ copii*
Sursa: Realizată de autor pe baza datelor oferite de ONU, 2016

Este clar că, după depășirea pragului de înlocuire a unei generații, viteza de variație a vârstei mediane este mai redusă.

2. Compararea a două modele de regresie. Modele cu interacțiune

Există deseori situația de a compara variația unei variabile dependente la nivelul a două grupe de populație, a două populații diferite sau pentru două perioade de timp diferite.

Pentru două eșantioane de volum n_1 și n_2, dacă se dispune de date statistice, se pot compara modelele de regresie care explică variația variabilei dependente.

Pentru exemplificare, considerăm cazul a două variabile între care există o dependență liniară. Cele două modele de regresie sunt date prin ecuațiile:

$$Y = \gamma_1 + \lambda_1 X + \varepsilon_1,\ i = \overline{1, n_1};$$

$$Y = \gamma_2 + \lambda_2 X + \varepsilon_2,\ i = \overline{1, n_2}.$$

Compararea celor două modele de regresie se realizează cu ajutorul unui model ANCOVA de forma:

$$Y = \alpha_0 + \alpha_1 D + \beta_1 X + \beta_2 (XD) + \varepsilon.$$

Variabila alternativă este:

- $D=0$, dacă $i=\overline{1,n_1}$;
- $D=1$, dacă $i=\overline{1,n_2}$.

Pentru acest model, regresia este:

$$E(Y/X,D)=\begin{cases}\alpha_0+\beta_1 X, & D=0\\ (\alpha_0+\alpha_1)+(\beta_1+\beta_2)X, & D=1\end{cases}$$

Prin testarea parametrilor modelului ANCOVA, se poate ajunge la următoarele 4 situaţii posibile:

- se acceptă ipotezele $H_0:\alpha_1=0$, $H_0:\beta_2=0$, rezultă $\gamma_1=\gamma_2$, $\lambda_1=\lambda_2$, adică între cele două modele de regresie nu există diferenţe semnificative;
- se acceptă ipoteza $H_0:\beta_2=0$ şi se respinge ipoteza $H_0:\alpha_1=0$, rezultă $\gamma_1\neq\gamma_2$, $\lambda_1=\lambda_2$, adică cele două regresii sunt paralele şi diferă doar prin ordonata la origine, având aceeaşi pantă;
- se acceptă ipoteza $H_0:\alpha_1=0$ şi se respinge ipoteza $H_0:\beta_2=0$, rezultă $\gamma_1=\gamma_2$, $\lambda_1\neq\lambda_2$, adică cele două regresii diferă prin viteza de variaţie, având doar aceeaşi ordonată la origine;
- se resping ipotezele $H_0:\alpha_1=0$, $H_0:\beta_2=0$, rezultă $\gamma_1\neq\gamma_2$, $\lambda_1\neq\lambda_2$, adică între cele două modele de regresie există diferenţe semnificative.

Ca interpretare, parametrul β_2 arată diferenţa de pantă sau de viteză de variaţie pentru cele două regresii aferente celor două grupe de unităţi din populaţie. În cazul în care parametrul este semnificativ statistic, acesta indică modificarea vitezei de variaţie a variabilei dependente care se înregistrează prin trecerea de la o grupă la alta.

Exemplu

Considerăm un eşantion de 102 ţări repartizate după vârsta mediană şi rata totală de fertilitate (date oferite de ONU, 2015). Grupăm ţările din eşantion în două categorii, în funcţie de variabila nominală *Continentul*, cu valorile *Asia* şi *Europa*.

În R, se construieşte variabila alternativă *D*:

- $D=0$, pentru ţările din Asia;
- $D=1$, pentru ţările din Europa.

Pentru a compara modelele de regresie aferente celor două grupe de ţări, se utilizează modelul ANCOVA:

$$Y=\alpha_0+\alpha_1 D+\beta_1 X+\beta_2(XD)+\varepsilon.$$

Rezultatele modelării în R se prezintă mai jos.

BCU IASI / CENTRAL UNIVERSITY LIBRARY

```
> summary(lm(Y~D1+X+X*D1))
Call:
lm(formula = Y ~ D1 + X + X * D1)
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  42,8114   1,1102      38,563   <2e-16 ***
D1 - Europe   5,6873   3,8035       1,495   0,138
X            -5,5287   0,4175     -13,241   <2e-16 ***
Europe: X     0,8482   2,3228       0,365   0,716
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 3,085 on 98 degrees of freedom
Multiple R-squared: 0,8505,     Adjusted R-squared: 0,8459
F-statistic: 185,9 on 3 and 98 DF,  p-value: < 2,2e-16
```

Testele Student ne arată că, pentru variabila dummy şi variabila *XD*, coeficienţii de regresie nu sunt semnificativ diferiţi de zero ($Pr(>|t|) > 0,05$, pentru ambii parametri). Concluzia este că între cele două modele nu există diferenţe semnificative. Din figura 6.10 se poate observa că cele două regresii pentru cele două grupe de ţări pe continente au viteze de variaţie egale (pantele dreptelor estimate nu diferă semnificativ).

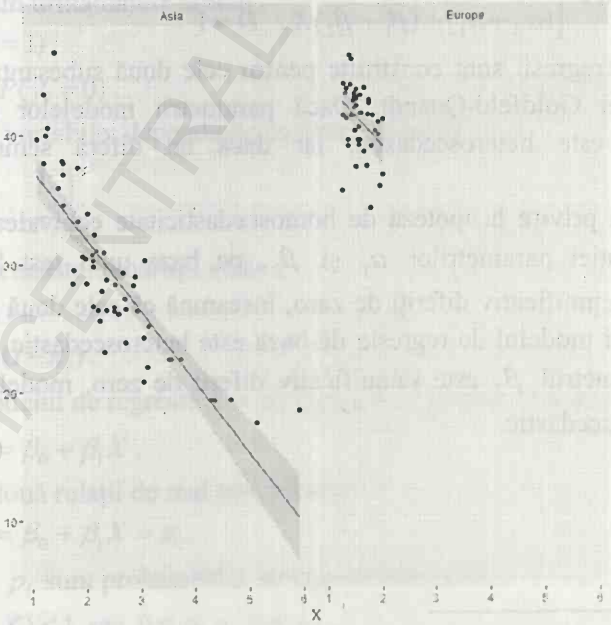


Figura 6.10. *Regresiile pentru cele două grupe de ţări pe continente*
Sursa: Realizată de autor pe baza datelor oferite de ONU, 2016

BCU IAŞI CENTRALĂ UNIVERSITY LIBRARY

3. Testarea ipotezei de homoscedasticitate a unui model de regresie

În testarea ipotezei de homoscedasticitate, o metodă mai facilă de testare, bazată pe testul Goldfeld-Quandt[1], o reprezintă utilizarea modelului ANCOVA cu o variabilă *dummy*. Variabila alternativă împarte seria de date în două părţi egale, în funcţie de o valoare critică x_0 a variabilei independente X, după omiterea unui număr de înregistrări din centrul seriei. Valoarea critică se stabileşte în funcţie de valorile variabilei independente şi este o valoare aproape de mediană sau poate coincide cu aceasta. Se obţin două seturi de date cu valori mici, respectiv cu valori mari, ale variabilei independente. Omiterea valorilor din centrul seriei empirice ordonate are scopul omogenizării celor două seturi de date.

Variabila *dummy* ia valorile:

- $D = 1$, dacă $x_i < x_0$;
- $D = 0$, în celălalt caz.

Pentru un model de regresie liniară simplă, testarea ipotezei de homoscedasticitate este echivalentă cu testarea semnificaţiei parametrilor corespunzători variabilei *dummy* din următorul model ANCOVA:

$$Y = \alpha_0 + \alpha_1 D + \beta_1 X + \beta_2 D \cdot X + \varepsilon .$$

Pentru acest model, mediile condiţionate sunt:

$$E(Y / X, D) = \begin{cases} \alpha_0 + \beta_1 X, & D = 0 \\ (\alpha_0 + \alpha_1) + (\beta_1 + \beta_2) X, & D = 1 \end{cases}$$

Cele două regresii sunt construite pentru cele două subeşantioane obţinute conform metodei Goldfeld-Quandt. Dacă parametrii modelelor diferă, atunci modelul iniţial este heteroscedastic, iar dacă nu diferă semnificativ, este homoscedastic.

Decizia cu privire la ipoteza de homoscedasticitate echivalează cu decizia asupra semnificaţiei parametrilor α_1 şi β_2, pe baza unui test Student. Dacă parametrii sunt semnificativ diferiţi de zero, înseamnă că cele două regresii diferă semnificativ, deci modelul de regresie de bază este heteroscedastic. Şi în cazul în care numai parametrul β_2 este semnificativ diferit de zero, modelul de regresie iniţial este heteroscedastic.

---

[1] Detaliile cu privire la ipoteza de homoscedasticitate şi la testarea acesteia vor fi prezentate în capitolul următor. De asemenea, exemplificarea utilizării modelului ANCOVA pentru testarea acestei ipoteze se va realiza tot în capitolul 7.

6.2. Modele de regresie cu variabile alternative dependente

În acest subcapitol, tratăm o serie de probleme care vizează modelele econometrice în care variabilele alternative apar ca variabile dependente şi prezentăm două dintre cele mai utilizate modele de acest tip: modelul *Logit* şi modelul *Probit*. În prima secţiune, se prezintă problemele metodologice specifice unui model de regresie în care variabila dependentă este o variabilă aleatoare alternativă.

6.2.1. Probleme ale estimării modelului cu variabilă alternativă dependentă

Considerăm modelul liniar de regresie simplă:

$Y = \beta_0 + \beta_1 X + \varepsilon$.

În acest model, variabila dependentă este o variabilă alternativă, care poate lua doar două valori:

- $Y = 0$, dacă nu se realizează o proprietate;
- $Y = 1$, dacă se realizează proprietatea cerută.

Fiind o variabilă dependentă, cele două valori ale variabilei *Y* apar în funcţie de valorile variabilei independente, nealternative, *X*. Ca variabilă aleatoare alternativă, *Y* urmează o lege de repartiţie Bernoulli.

Considerăm următoarele notaţii:

$\pi_i = P(y_i = 1)$;

$(1 - \pi_i) = P(y_i = 0)$.

Repartiţia variabilei dependente se prezintă astfel:

$$Y : \begin{pmatrix} 0 & 1 \\ 1 - \pi_i \cdots & \pi_i \end{pmatrix}.$$

Parametrii acestei repartiţii sunt:

$E(Y) = \pi_i$,

$V(Y) = \pi_i(1 - \pi_i)$.

Pentru modelul de regresie $Y = \beta_0 + \beta_1 X + \varepsilon$, media condiţionată este:

$E(Y / X) = \beta_0 + \beta_1 X$.

Din cele două relaţii de mai sus, rezultă:

$E(Y / X) = \beta_0 + \beta_1 X = \pi_i$.

Deoarece p_i sunt probabilităţi, se impune restricţia:

$0 \le E(Y / X) \le 1$ sau $0 \le \beta_0 + \beta_1 X \le 1$.

BCU IASI / CENTRAL UNIVERSITY LIBRARY

Ţinând cont de legea de repartiţie a variabilei dependente, pentru variabila reziduală $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$, se poate scrie următoarea repartiţie:

$$\varepsilon_i : \begin{pmatrix} -\beta_0 - \beta_1 x_i & 1 - \beta_0 - \beta_1 x_i \\ 1 - \pi_i & \pi_i \end{pmatrix},$$

deoarece variabila dependentă ia valorile 0 şi 1.

Pentru repartiţia erorilor, caracteristicile numerice sunt:

- media sau speranţa matematică:
  $E(\varepsilon_i) = (1-\pi_i)(-\beta_0 - \beta_1 x_i) + \pi_i(1 - \beta_0 - \beta_1 x_i) = 0$;
- dispersia sau varianţa:
  $V(\varepsilon_i) = E(\varepsilon_i^2) = (1-\pi_i)(-\beta_0 - \beta_1 x_i)^2 + \pi_i(1 - \beta_0 - \beta_1 x_i)^2 = \pi_i(1-\pi_i)$.

Observaţii

Pe baza rezultatelor de mai sus, se pot sintetiza următoarele observaţii asupra modelului de regresie liniară cu variabilă dependentă alternativă:

- curba de regresie impune restricţia $0 \le \beta_0 + \beta_1 x_i \le 1$. În practică însă, nu se poate garanta că are loc relaţia $0 \le \hat{Y} \le 1$. De asemenea, această restricţie impune ca probabilitatea π_i să crească liniar, în funcţie de valorile x_i.
- sunt încălcate două ipoteze clasice ale modelului de regresie: ipoteza de normalitate a erorilor (erorile nu urmează o lege de repartiţie normală, ci pot lua doar două valori, conform tabloului de mai sus) şi ipoteza de homoscedasticitate ($V(\varepsilon_i) = \pi_i(1-\pi_i) = \sigma_i^2 \neq const$).

Înlăturarea acestor inconveniente se poate realiza prin construirea unor modele derivate precum: modelul *Logit*, modelul *Probit*, modelul *Tobit*.

6.2.2. Modelul Logit

Acest model are la bază ideea de a considera, pentru media condiţionată $E(Y/X = x_i) = \pi_i$, o relaţie care are la bază repartiţia logistică.

În modelul *Logit*, pentru media condiţionată, se utilizează relaţia:

$$E(Y/X = x_i) = \pi_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \text{ sau}$$

$$\pi_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} = \frac{1}{1 + e^{-z_i}}.$$

Cu alte cuvinte, regresia este dată de funcţia $\pi_i = F(z_i)$, unde F este funcţia de repartiţie specifică repartiţiei logistice.

Aplicând funcţia inversă, se obţine:

BCU ASSOCIATED UNIVERSITY LIBRARY

$z_i = \ln\frac{\pi_i}{1-\pi_i}$, iar

$z_i = \beta_0 + \beta_1 x_i$.

Pe baza relaţiilor de mai sus, se defineşte modelul logistic:

$$L_i = \ln\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_i + \varepsilon_i .$$

Observaţii

Acest model înlătură restricţia ca media condiţionată să ia valori în intervalul *[0,1]*. Într-adevăr, dacă:

$\pi_i = 0 \Rightarrow L_i \to -\infty$;

$\pi_i = 1 \Rightarrow L_i \to +\infty$.

Dacă se consideră un eşantion de volum mare, variabila eroare urmează aproximativ o lege normală cu următorii parametri:

$$\varepsilon_i \sim N(0, \frac{1}{N_i \pi_i (1-\pi_i)}),$$

unde N_i este numărul de unităţi din populaţie care au valoarea x_i pentru variabila independentă. Probabilităţile π_i pot fi estimate cu ajutorul relaţiei:

$$p_i = \frac{n_i}{N_i},$$

unde n_i este numărul de unităţi din cele N_i care îndeplinesc proprietatea pentru care *Y=1*.

În modelul logistic, termenul $\frac{\pi_i}{1-\pi_i}$ se numeşte *raport de şanse (odds)*, iar interpretarea parametrilor este următoarea:

- β_0 reprezintă logaritmul raportului de şanse, în condiţiile în care variabila independentă este nulă (*X=0*). Raportul de şanse se obţine prin funcţia inversă, adică: *odds* = e^{β_0}, dacă *X=0*;
- β_1 reprezintă modificarea relativă a raportului de şanse, dacă variabila independentă se modifică absolut cu o unitate:

$$\beta_1 = d \ln\frac{\pi_i}{1-\pi_i}\Big|_{dX=1}.$$

Utilizând proprietatea funcţiei logaritm, variaţia relativă a raportului de şanse este tocmai logaritmul unui raport între două valori ale raportului de şansă: cea care s-a înregistrat după ce are loc o variaţie cu o unitate a variabilei

independente și cea de dinainte de această variație. În literatura de specialitate, acest raport dintre două valori ale raportului de șanse poarte numele de *odds ratio.*

O interpretare mai facilă a rezultatelor obținute cu ajutorul modelului logistic se obține prin funcția inversă aplicată relației $\beta_1 = d \ln \frac{\pi_i}{1-\pi_i}\Big|_{dX=1}$. Rezultă:

Odds ratio $= e^{\beta_1}$.

Această valoare arată de câte ori se modifică raportul de șanse, dacă variabila independentă se modifică cu o unitate.

Etapele estimării modelului Logit

Pentru estimarea parametrilor unui model *Logit*, se parcurg următoarele etape:

1. Estimarea parametrilor π_i

Pentru fiecare valoare x_i a variabilei independente, se determină valorile estimate $p_i = \frac{n_i}{N_i}$ pentru parametrii π_i.

2. Calculul estimațiilor valorilor L_i

Se calculează valorile $l_i = \ln \frac{p_i}{1-p_i}$, ca estimații pentru L_i.

3. Construirea modelului de regresie

Pe baza estimațiilor obținute în etapa anterioară, se estimează modelul de regresie:

$L = \beta_0 + \beta_1 X + \varepsilon$.

4. Corectarea heteroscedasticității

Deoarece modelul *Logit* este heteroscedastic, se corectează cu ajutorul metodei celor mai mici pătrate generalizate, prin ponderare cu ajutorul estimațiilor dispersiilor erorilor:

$$s_i^2 = \frac{1}{N_i p_i (1-p_i)}.$$

Modelul ponderat este:

$$\frac{L_i}{s_i} = \frac{\beta_0}{s_i} + \beta_1 \frac{x_i}{s_i} + \frac{\varepsilon_i}{s_i}.$$

Acest model poate fi transformat într-un model liniar și se estimează prin metoda celor mai mici pătrate.

BCU "CAROL I" UNIVERSITY LIBRARY

Observaţie

Modelul *Logit* se poate generaliza pentru un anumit număr de predictori sau de variabile independente. Ecuaţia de regresie este de forma:

$$E(Y/X_j)=\pi_i=\frac{1}{1+e^{-(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+...+\beta_p x_{pi})}}.$$

Exemplul 1

Se consideră un set de date convenţionale privitoare la disponibilitatea unui angajat de a accepta orice condiţii la locul de muncă, în funcţie de nivelul de educaţie (ani de studiu). Datele sunt prezentate în tabelul 6.1.

Tabelul 6.1. *Date privind disponibilitatea angajaţilor*

Educaţie (ani)	Persoane intervievate	Persoane care acceptă orice condiţii
4.00	25.00	22.00
8.00	30.00	27.00
10.00	20.00	16.00
12.00	40.00	30.00
14.00	10.00	7.00
15.00	20.00	15.00
16.00	30.00	15.00
17.00	15.00	6.00
18.00	25.00	5.00
20.00	30.00	5.00
22.00	10.00	2.00

Sursa: Realizat de autor (date convenţionale)

Pentru datele din tabelul de mai sus, în R, s-au obţinut următoarele rezultate ale modelării cu ajutorul modelului *Logit*:

```
> summary(glm(Y~X, family = "binomial"))

Call:
glm(formula = Y ~ X, family = "binomial")
Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  4,20997    0,58620      7,182      6,88e-13 ***
X            -0,27018   0,03816      -7,080     1,44e-12 ***
---
Signif. codes:  0 '***'  0,001 '**'  0,01 '*'  0,05 '.'  0,1 ' '  1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 345,52  on 254  degrees of freedom
Residual deviance: 271,91  on 253  degrees of freedom
```

AIC: 275,91
Number of Fisher Scoring iterations: 4

Modelul *Logit* estimat are ecuația:

$L = 4{,}21 - 0{,}27X$ sau

$$\ln\frac{p}{1-p} = 4{,}21 - 0{,}27X\,.$$

Pentru a obține valorile coeficienților prin funcția exponențială, în R, se utilizează comanda:

```
> exp(coef (glm(Y~X, family = "binomial")))
```

```
(Intercept)          X
67,3546637   0,7632456
```

Curba de regresie se obține în R astfel:

```
> plot(Y~X)
> curve(exp(4.21-0.27*x)/(1+exp(4.21-0.27*x)), add=T)
```

Figura 6.11. *Regresia logistică pentru variabilele disponibilitate și nivel de educație*
Sursa: Realizată de autor pe baza datelor din tabelul 6.1

Interpretare

- valoarea $e^{b_0} = 67{,}35$ este valoarea raportului de șanse pentru un angajat fără studii ($X=0$). Cu alte cuvinte: $\frac{p}{1-p} = 67{,}35$, dacă $X=0$. Adică, un angajat fără studii are probabilitatea de 67,5 ori mai mare de a accepta condițiile de lucru față de situația de a nu le accepta.

BCU IASI CENTRAL UNIVERSITY LIBRARY

- valoarea $e^{b_1} = 0,76$ arată că raportul de şanse creşte de 0,76 ori (adică scade cu 34%) dacă nivelul studiilor creşte cu un an. Estimaţia $b_1 = 0,27$ este negativă şi arată existenţa unei legături negative între variabile.

Modelul ne permite estimarea unei probabilităţi π ca o persoană cu un anumit nivel al studiilor x_i să accepte orice tip de condiţie la locul de muncă.

De exemplu, pentru o persoană cu o educaţie de 15 ani de studii, probabilitatea de a accepta orice condiţii de muncă se obţine astfel:

$$\ln\frac{p}{1-p} = 4,21 - 0,27 \cdot 15 = 0,15.$$

Rezultă:

$$\frac{p}{1-p} = 1,161 \text{ şi } p = 0,53.$$

Exemplul 2

Considerăm rezultatele unei anchete pe un eşantion de 513 femei, realizată la maternitatea Cuza Vodă din Iaşi în cadrul proiectului internaţional *Précarité, renoncement aux soins et périnatalité: Une comparaison franco-roumaine*, finanţat de Universitatea din Poitiers.

În cadrul anchetei, au fost înregistrate date privind renunţarea mamelor la consultaţii în timpul sarcinii (variabila dependentă, cu 2 valori: Da - valoarea 1; Nu - valoarea 0). Cu ajutorul unui model logistic se poate estima probabilitatea de a renunţa la asistenţa medicală de sănătate în timpul sarcinii în funcţie de vârsta mamei (variabila X_1, ani) şi numărul de copii născuţi deja (variabila X_2, persoane).

Modelarea în R presupune aplicarea comenzii:

```
> summary(glm(Y~X1+X2, family = "binomial"))
```

Rezultatele sunt prezentate mai jos.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2,20723	0,51270	4,305	1,67e-05 ***
X1	-0,14747	0,02095	-7,039	1,94e-12 ***
X2	0,69952	0,09391	7,448	9,45e-14 ***

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 585,05 on 512 degrees of freedom

Residual deviance: 495,04 on 510 degrees of freedom

AIC: 501,04

Modelul estimat este de forma:

$$\ln\frac{p}{1-p} = 2{,}21 - 0{,}14X_1 + 0{,}69X_2 .$$

Testul Student arată că ambele variabile independente sunt semnificative statistic, X_1 are un impact negativ, iar X_2 are un impact pozitiv asupra variabilei dependente.

Estimațiile coeficienților prin funcția exponențială se obțin mai jos:

```
> exp(coef(glm(y~X1+X2, family = "binomial")))
```

(Intercept)	X1	X2
9,0905159	0,8628909	2,0127776

Interpretare

- $e^{b_0} = 9{,}09$ este valoarea raportului de șanse pentru o persoană în cazul în care variabilele independente sunt nule (este o valoare care nu are sens practic);
- valoarea $e^{b_1} = 0{,}86$ arată că raportul de șanse crește de 0,98 ori (adică scade cu 2%), dacă nivelul vârstei crește cu un an, iar numărul de copii născuți rămâne constant;
- valoarea $e^{b_2} = 2{,}01$ arată că raportul de șanse crește de 2,01 ori (adică crește cu 201%), dacă numărul de copii născuți de o femeie crește cu o unitate, iar nivelul vârstei rămâne constant. Cu alte cuvinte, probabilitatea de a renunța la îngrijirea sănătății în perioada de sarcină scade în condițiile creșterii vârstei și crește în condițiile în care crește numărul de copii născuți deja.

6.2.3. Modelul Probit

În acest model, se utilizează funcția de repartiție pentru o repartiție normală:

$$F(z) = \int_{-\infty}^{z} f(t)dt \text{, unde } f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} .$$

După o schemă similară modelului *Logit*, considerăm că probabilitățile π_i sunt date de funcția de repartiție a unei repartiții normale: $\pi_i = F(z_i)$, unde:

$$Z = \beta_0 + \beta_1 X .$$

Pentru determinarea funcției de repartiție, se consideră o valoare critică z_i^*, în funcție de care se realizează sau nu proprietatea cerută pentru variabila dependentă:

$Y = 0$, dacă $z_i < z_i^*$;

BCU IASI / CENTRAL UNIVERSITY LIBRARY

$Y = 1$, dacă $z_i \geq z_i^*$.

Astfel, $\pi_i = P(Y=1) = P(z_i^* < z_i) = F(z_i) = \int_{-\infty}^{z_i} f(t)dt = \int_{-\infty}^{\beta_0+\beta_1 x_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$.

Prin aplicarea funcţiei inverse, din relaţia $\pi_i = F(z_i)$ sau $\pi_i = F(\beta_0 + \beta_1 x_i)$ se obţine:

$z_i = \beta_0 + \beta_1 x_i = F^{-1}(\pi_i)$.

Relaţia dintre probabilităţi şi răspunsul prin funcţia inversă $F^{-1}(\pi_i)$ este prezentată în figura 6.12.

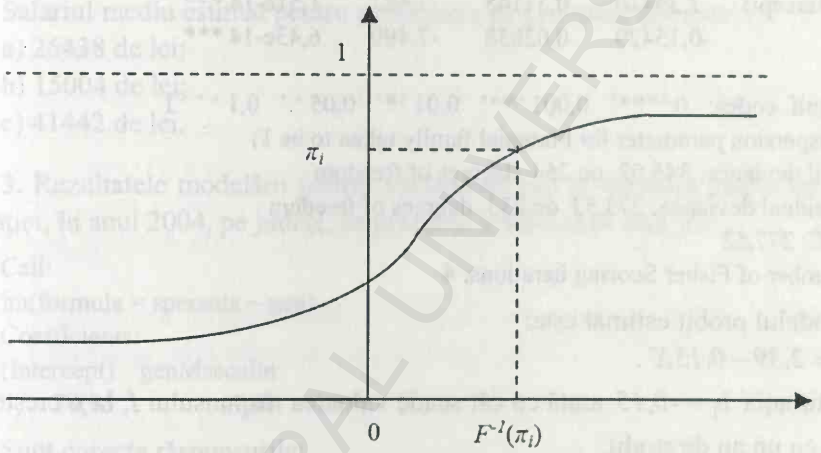


Figura 6.12. *Modelul Probit*

Pe baza relaţiilor de mai sus, se defineşte modelul Probit:

$I_i = F^{-1}(\pi_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Observaţie

Deoarece valorile funcţiei $F^{-1}(\pi_i)$ pot fi şi negative, în practică, estimarea parametrilor modelului se realizează pe baza unui artificiu, prin sumarea cu 5 a fiecărei valori I_i.

Etapele estimării modelului Probit

- se calculează valorile $p_i = \frac{n_i}{N_i}$;
- pe baza repartiţiei normale, se obţin valorile $I_i = F^{-1}(p_i)$. Pentru modelul *Probit*, la aceste valori se sumează valoarea 5;

BCU "CAROL I" CENTRAL UNIVERSITY LIBRARY

- se construieşte şi se estimează modelul de regresie $I_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ cu ajutorul valorilor determinate anterior.

Exemplu

Se utilizează aceleaşi date ca în cazul modelului *Logit*. În R, rezultatele modelării sunt prezentate mai jos.

```
> summary(glm(Y~X, family = binomial(link="probit")))

Call:
glm(formula = Y ~ X, family = binomial(link = "probit"))
Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  2,39410    0,31145      7,687      1,51e-14 ***
X            -0,15429   0,02058      -7,499     6,43e-14 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 345,52  on 254  degrees of freedom
Residual deviance: 273,52  on 253  degrees of freedom
AIC: 277,52
Number of Fisher Scoring iterations: 4
```

Modelul probit estimat este:

$I = 2,39 - 0,15X$.

Estimaţia $b_1 = -0,15$ arată cu cât scade valoarea răspunsului *I*, la o creştere a educaţiei cu un an de studii.

Pe baza modelului, se pot realiza estimaţii ale probabilităţii răspunsului pentru o anumită valoare a variabilei independente *X*. De exemplu, pentru un salariat cu 15 ani de studii, rezultă:

$I = 2,39 - 0,15 \cdot 15 = 0,14$.

Din relaţia $F^{-1}(w) = 0,14$, rezultă $p = 0,55$, adică un salariat cu 15 ani de studii are 55 de şanse dintr-o sută de a accepta orice condiţii de lucru la locul de muncă.

6.3. Teste şi exerciţii

Teste grilă

1. Analiza influenţei nivelului educaţiei (primar, mediu, superior) asupra venitului, pentru un eşantion de angajaţi, se poate realiza cu ajutorul unui model:
a) ANOVA cu 3 variabile *dummy*;
b) econometric, pentru serii de timp;

BCU "ASCENTRAL UNIVERSITY LIBRARY

c) ANOVA cu 2 variabile *dummy*;
d) ANCOVA;
e) econometric, pentru serii de moment.

2. Rezultatele modelării pentru variabilele gen şi salariu (lei), la nivelul unui eşantion de firme, în anul 2005, se prezintă în tabelul de mai jos.

Call:
lm(formula = salary ~ gender)
Coefficients:
(Intercept) genderMale
26438 15004

Salariul mediu estimat pentru persoanele de gen masculin este:
a) 26438 de lei;
b) 15004 de lei;
c) 41442 de lei.

3. Rezultatele modelării pentru variabilele gen şi speranţa medie de viaţă a populaţiei, în anul 2004, pe judeţe, se prezintă în tabelul de mai jos.

Call:
lm(formula = speranta ~ gen)
Coefficients:
(Intercept) genMasculin
74,95 -7,41

Sunt corecte răspunsurile:
a) ecuaţia modelului estimat este: $Y = 74,95 - 7,41D$;
b) estimaţia $a_0 = 74,95$ *de ani* este speranţa de viaţă medie masculină estimată la nivelul unui judeţ al României;
c) estimaţia $a_0 + a_1 = 74,95 - 7,41 = 67,54$ *de ani* este speranţa de viaţă medie masculină estimată la nivelul unui judeţ al României;
d) $a_1 = -7,41$ *de ani* este estimaţia diferenţei dintre speranţa medie de viaţă masculină şi cea feminină la nivelul unui judeţ.

4. Rezultatele modelării pentru variabilele gen şi salariu (lei), la nivelul unui eşantion de firme, în anul 2005, se prezintă în tabelul de mai jos.

Call:
lm(formula = salary ~ gender)
Coefficients:

BCU CENTRAL UNIVERSITY LIBRARY

(Intercept) genderMale
26438 15004

Diferența dintre salariul mediu estimat al persoanele de gen masculin și cel al persoanelor de gen feminin este:

a) 26438 de lei;

b) 15004 de lei;

c) 41442 de lei.

5. Rezultatele modelării pentru variabilele gen și speranța medie de viață a populației, în anul 2004, pe județe, se prezintă în tabelul de mai jos.

Call:
lm(formula = speranta ~ gen)
Coefficients:
(Intercept) genMasculin
74,95 -7,41

Au loc răspunsurile:

a) regresia estimată pentru populația de gen masculin este: $Y = 67,54$;

b) în medie, o femeie are o speranță de viață mai scurtă decât un bărbat, cu 7,41 ani;

c) regresia estimată pentru populația de gen feminin este: $Y = 74,95$;

d) dreapta de regresie estimată pentru fiecare grup are panta egală cu 0.

6. Rezultatele modelării pentru variabilele gen, nivelul de educație (ani) și nivelul salariului ($), pentru un eșantion de angajați, se prezintă în tabelul de mai jos.

Call:
lm(formula = speranta ~ gen+educ)
Coefficients:
(Intercept) genMasculin educ
-15925 8500 3400

Sunt corecte afirmațiile:

a) $a_0 = -15925$ $ este nivelul mediu estimat al salariului pentru persoanele de gen feminin, în condițiile în care nivelul studiilor este $X=0$;

b) $a_1 = 8500$ $ este diferența dintre salariul mediu al bărbaților și cel al femeilor;

c) $b = 3400$ $ este creșterea salariul mediu al unui angajat de gen masculin la o creștere a nivelului de educație cu un an;

BCU "CAROL I" CENTRAL UNIVERSITY LIBRARY

d) un salariat de gen feminin câştigă, în medie, cu *8500*$ mai mult decât un angajat de gen masculin.

7. Rezultatele modelării pentru variabilele gen, nivelul de educaţie (ani) şi nivelul salariului ($), pentru un eşantion de angajaţi, se prezintă în tabelul de mai jos.

Call:
lm(formula = speranta ~ gen+educ)
Coefficients:

(Intercept)	genMasculin	educ
-15925	8500	3400

Au loc rezultatele:

a) regresia pentru persoanele de gen masculin este: $Y_X = -7425 + 3400X$;

b) regresiile pentru cele două grupe de salariaţi au aceeaşi pantă;

c) regresia pentru persoanele de gen masculin este:
$Y_X = -15925 + 8500D + 3400X$;

d) regresiile pentru cele două grupe de salariaţi au aceeaşi ordonată la origine;

8. Rezultatele modelării pentru variabilele gen, timpul lucrat în companie (*jobtime*, luni) şi nivelul salariului ($), pentru un eşantion de angajaţi, se prezintă în tabelul de mai jos.

Call:
lm(formula = salariu ~ gen + jobtime)
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18981,25	5993,27	3,167	0,00165**
genMale	14873,25	1490,83	9,976	< 2e-16***
jobtime	92,84	73,29	1,267	0,20589

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Au loc rezultatele:

a) regresia pentru persoanele de gen feminin este: $Y_X = 18981{,}25 + 92{,}84X$;

b) variabilele semnificative din model sunt *genul* şi *jobtime*;

c) variabila *jobtime* ar trebui eliminată din model.

BCU IASI / CENTRAL UNIVERSITY LIBRARY

Exerciții

1. Se consideră o variabilă dependentă cantitativă Y și o variabilă calitativă cu două valori. Pentru această variabilă, se construiesc 2 variabile alternative, conform tabloului de mai jos:

Condiție	D_1
A	1
în rest	0

Condiție	D_2
A^c	1
în rest	0

Se poate modela legătura dintre variabile cu ajutorul ecuației de mai jos:
$Y = \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + \varepsilon$?

Ce probleme apar și ce soluții se pot admite?

2. Rezultatele modelării pentru variabilele gen, experiența anterioară (*prevexp,* luni) și nivelul salariului ($), pentru un eșantion de angajați, se prezintă în tabelul de mai jos.

Call:
lm(formula = salariu ~ gen + prevexp + gen * prevexp)
Coefficients:

(Intercept)	genMale	prevexp	genMasculin* prevexp
28417,4	17034,1	-22,84	-13,09

Se cere:

a) să se scrie ecuația de regresie estimată;

b) să se interpreteze semnificația parametrilor;

c) să se scrie regresiile pentru cele două grupe de angajați determinate de variabila gen.

3. Rezultatele modelării pentru variabilele gen, experiența anterioară (*prevexp,* luni) și nivelul salariului ($), pentru un eșantion de angajați, se prezintă în tabelul de mai jos.

Call:
lm(formula = salariu ~ gen + prevexp + gen * prevexp)
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28417,41	1484,85	19,138	< 2e-16***
genMasculin	17034,12	2013,37	8,460	3,85e-16***
prevexp	-22,84	11,46	-1,993	0,0469*
genMasculin*prevexp	-13,09	14,39	-0,910	0,3635

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

BCU IASI / CENTRAL UNIVERSITY LIBRARY

Residual standard error: 15290 on 446 degrees of freedom
Multiple R-squared: 0,2216, Adjusted R-squared: 0,2164
F-statistic: 42,33 on 3 and 446 DF, p-value: < 2,2e-16

Se cere:

a) să se testeze coeficienții de regresie, cu o încredere de 95%, și să se stabilească ce variabile au impact semnificativ asupra salariului;

b) să se testeze modelul de regresie, cu o încredere de 95%;

c) interpretați valoarea estimată a coeficientului de determinație și a coeficientului de determinație ajustat.

4. Rezultatele modelării pentru variabilele gen, experiența în muncă (*exp*, luni), departament (management, depozit, contabilitate) și nivelul salariului (S), pentru un eșantion de angajați, se prezintă în tabelul de mai jos.

Call:
lm(formula = salariu ~ gen + exp + depart)
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26121,891	820,140	31,851	< 2e-16***
genMasculin	7590,069	997,879	7,606	1,69e-13***
exp	15,819	4,931	3,208	0,00143**
departDepozit	1942,799	2233,793	0,870	0,38492
departManag	32397,286	1247,174	25,977	< 2e-16***

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 9621 on 445 degrees of freedom
Multiple R-squared: 0,6926, Adjusted R-squared: 0,6898
F-statistic: 250,7 on 4 and 445 DF, p-value: < 2,2e-16

Se cere:

a) să se scrie ecuația de regresie estimată;

b) să se interpreteze semnificația parametrilor;

c) să se testeze coeficienții de regresie, cu o încredere de 95%, și să se stabilească ce variabile au impact semnificativ asupra salariului;

d) să se scrie regresiile pentru toate grupele de angajați determinate de variabilele calitative;

e) să se testeze modelul de regresie, cu o încredere de 95%;

f) să se analizeze ideea de a reduce numărul de grupe pentru variabila calitativă *departament* și ideea reconstruirii modelului de regresie;

g) să se estimeze cu cât va crește, în medie, salariul unui angajat de gen masculin din departamentul de contabilitate, dacă experiența în muncă crește cu 12 luni;

e) să se estimeze cu cât ar trebui să crească experiența unei persoane de gen feminin din departamentul de management pentru a înregistra o majorare de salariu cu 1000$.

5. Rezultatele modelării pentru variabilele gen, experiența în muncă (*exp*, luni), departament (management, depozit, contabilitate) și nivelul salariului ($), pentru un eșantion de angajați, se prezintă în tabelul de mai jos.

Call:
lm(formula = salariu ~ gen + exp + depart + gen*exp)
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26252,284	938,834	27,963	< 2e-16***
genMasculin	7338,798	1329,642	5,519	5,79e-08***
exp	17,328	7,221	2,400	0,0168*
departDepozit	1670,575	2429,828	0,688	0,4921
departManag	32404,886	1248,744	25,950	< 2e-16***
genMasculin*exp	2,828	9,876	0,286	0,7748

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 9631 on 444 degrees of freedom
Multiple R-squared: 0,6927, Adjusted R-squared: 0,6892
F-statistic: 200,1 on 5 and 444 DF, p-value: < 2,2e-16

Se cere să se răspundă la toate cerințele de la exercițiul anterior. Să se comenteze diferențele între cele două modele.

BCU IASI/CENTRALA UNIVERSITARA LIBRARY

Capitolul 7

VERIFICAREA IPOTEZELOR MODELULUI DE REGRESIE

Modelarea econometrică se realizează în anumite condiţii sau cu respectarea unui set de restricţii care se numesc *ipoteze ale modelului de regresie*. Calitatea estimării parametrilor modelului de regresie depinde de îndeplinirea a două clase de ipoteze: ipoteze asupra componentei aleatoare sau asupra variabilei eroare şi ipoteze asupra componentei deterministe sau asupra variabilelor independente.

Ipotezele asupra componentei aleatoare sunt: media erorilor este nulă, homoscedasticitatea, normalitatea şi necorelarea erorilor. Formal, aceste ipoteze se scriu astfel:

- $E(\varepsilon_i) = 0$, media erorilor este nulă;
- $V(\varepsilon_i) = \sigma^2$, ipoteza de homoscedasticitate;
- $\varepsilon_i \sim N(0, \sigma^2)$, ipoteza de normalitate;
- $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, ipoteza de necorelare sau de independenţă a erorilor.

Ipotezele asupra componentei deterministe sunt:

- variabilele independente sunt non-aleatoare, deterministe;
- variabilele independente sunt necoliniare;
- variabilele independente şi variabila eroare sunt necorelate, $\text{cov}(X_j, \varepsilon) = 0$.

Nerespectarea acestor ipoteze determină modificarea proprietăţilor estimatorilor parametrilor modelului de regresie şi ridică probleme importante în realizarea demersului cercetării econometrice.

BCU IASI / CENTRAL UNIVERSITY LIBRARY

7.1. Ipoteze asupra erorilor

Pentru testarea ipotezelor cu privire la componenta aleatoare, se va aborda un demers care presupune parcurgerea următoarelor etape: definirea ipotezei, stabilirea efectelor încălcării ipotezei, testarea ipotezei pe un set de date statistice și corectarea modelului, în condițiile în care ipoteza este încălcată.

7.1.1. Media variabilei reziduale este egală cu zero

Formal, ipoteza se prezintă prin relația: $E(\varepsilon_i) = 0$.

1. Definirea ipotezei

Potrivit acestei ipoteze, restricția modelării econometrice este ca toți ceilalți factori neincluși în model, reprezentați de variabila reziduală, precum și erorile determinate de metoda statistică să nu afecteze sistematic media variabilei dependente Y.

Ipoteza $E(\varepsilon_i) = 0$ este echivalentă cu condiția:

$E(Y/X) = \beta_0 + \beta_1 X$.

2. Efectele încălcării ipotezei

Dacă media variabilei reziduale nu este egală cu zero, atunci se modifică proprietățile estimatorilor parametrilor modelului de regresie. Avem două situații: când media variabilei reziduale este constantă și când aceasta nu este constantă.

a. $E(\varepsilon_i) = \mu = cst.$

Considerăm modelul de regresie liniară simplă:

$Y = \beta_0 + \beta_1 X + \varepsilon$.

Acesta se mai poate scrie:

$Y = \beta_0 + \mu + \beta_1 X + \varepsilon - \mu = \beta_0^* + \beta_1 X + \varepsilon^*$,

unde $\beta_0^* = \beta_0 + \mu$, $\varepsilon^* = \varepsilon - \mu$.

Pentru ultimul model obținut, $Y = \beta_0^* + \beta_1 X + \varepsilon^*$, este îndeplinită ipoteza $E(\varepsilon_i^*) = 0$, însă se poate arăta că parametrul β_1 este estimat nedeplasat de estimatorul $\hat{\beta}_1$, iar parametrul β_0 este estimat deplasat de estimatorul $\hat{\beta}_0^*$. Astfel:

$E(\hat{\beta}_0^*) = \beta_0 + \mu$.

b. $E(\varepsilon_i) = \mu_i$

În acest caz, modelul de regresie se poate scrie:

$y_i = \beta_0 + \mu_i + \beta_1 x_i + \varepsilon_i - \mu_i = \beta_0^* + \beta_1 x_i + \varepsilon_i^*$

BCU "CAROL I" UNIVERSITY LIBRARY

şi se poate demonstra că parametrul β_1 este estimat deplasat de estimatorul $\hat{\beta}_1$.

Într-adevăr, dacă considerăm următoarea relaţie:

$$\hat{\beta}_1 = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} = \sum_i \frac{nx_i - \sum_i x_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} \cdot y_i ,$$

rezultă că media estimatorului $\hat{\beta}_1$ va fi:

$$E(\hat{\beta}_1) = \sum_i \frac{nx_i - \sum_i x_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} \cdot E(y_i) , \text{ iar}$$

$$E(y_i) = \beta_0 + \mu_i + \beta_1 x_i .$$

Rezultă,

$$E(\hat{\beta}_1) = \beta_1 + \frac{n\sum_i x_i \mu_i - \sum_i x_i \sum_i \mu_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} ,$$

ceea ce indică un estimator deplasat.

În concluzie, dacă ipoteza $E(\varepsilon_i) = 0$ este încălcată, estimarea parametrilor modelului se realizează cu o eroare sistematică. Este vorba despre o deplasare de care suferă fie estimarea parametrului β_0, fie estimarea parametrului β_1.

3. Testarea ipotezei cu privire la media erorilor

Verificarea acestei ipoteze se realizează pe un set de date, de obicei, de la nivelul unui eşantion.

Etapele testării sunt următoarele:

- se estimează un model de regresie liniară simplă, fără a ţine cont de ipoteza cu privire la media erorilor;
- se determină erorile estimate, ca diferenţă între valorile variabilei dependente observate şi cele calculate pe baza modelului estimat. Erorile estimate sunt de forma $e_i = y_i - b_0 - b_1 x_i$;
- se realizează un test Student cu privire la media erorilor, în care ipoteza nulă este: $H_0: \mu_\varepsilon = 0$;
- rezultatul testării, pentru un prag de semnificaţie stabilit, ne arată dacă este încălcată sau nu ipoteza $E(\varepsilon_i) = 0$.

BCU "MIHAIL KOGĂLNICEANU" UNIVERSITY LIBRARY

Testul Student pentru testarea ipotezei $E(\varepsilon)=0$

Formularea ipotezelor

$H_0: E(\varepsilon)=0;$

$H_1: E(\varepsilon)\neq 0.$

Alegerea pragului de semnificație

$\alpha=0{,}05.$

Alegerea testului

Se utilizează statistica Student:

$$t=\frac{\hat{E}(\varepsilon)-E(\varepsilon)}{\sqrt{\hat{V}(\hat{E}(\varepsilon))}}.$$

În condițiile acceptării ipotezei nule, statistica Student este:

$$t=\frac{\hat{E}(\varepsilon)}{\sqrt{\hat{V}(\hat{E}(\varepsilon))}}.$$

Valoarea teoretică a testului

Din tabela Student, se citește valoarea: $t_{\alpha/2;n-1}$.

Valoarea calculată a testului

$$t_{calc}=\frac{E(e_i)}{s_{\hat{E}(\varepsilon)}}.$$

Decizia

Se compară valoarea calculată a testului cu valoarea teoretică.

4. Corectarea modelului

Dacă ipoteza cu privire la media erorilor este încălcată, soluția este corectarea modelului inițial cu ajutorul estimației mediei erorilor calculate la nivelul setului de date disponibile. Astfel, dacă ceilalți factori, neincluși în model, induc o deplasare sau o influență sistematică asupra mediei variabilei dependente, atunci valorile variabilei dependente pot fi corectate cu această valoare.

Modelul corectat va fi de forma:

$y_i^{\bullet}=\beta_0+\beta_1 x_i+u_i$, unde

$y_i^{\bullet}=y_i-E(\varepsilon_i).$

Observații

- demersul testării este același pentru orice tip de model econometric utilizat (liniar sau neliniar, simplu sau multiplu);

BCU "CAROL I" CENTRAL UNIVERSITY LIBRARY

- încălcarea ipotezei are efect asupra parametrului β_0, care frecvent în practică nu are o semnificație concretă sau nu este important pentru rezultate;
- păstrarea parametrului β_0 în model, chiar dacă nu este semnificativ statistic, conduce în final la o soluție convenabilă pentru verificarea ipotezei $E(\varepsilon)=0$.

Exemplu

Pentru exemplificare, considerăm datele prezentate la capitolul 3, secțiunea 3.7, privind populația școlară din învățământul superior (Y) și Produsul Intern Brut (X), pe județe ale României, în anul 2015.

În R, s-au obținut rezultatele:

```
> lm(Y~X)

Call:
lm(formula = Y ~ X)
Coefficients:
(Intercept)        X
 8273,1445   0,5283
```

Modelul estimat este:

$Y_X = 8273{,}144 + 0{,}528X$.

Pe baza ecuației modelului estimat, se obțin estimațiile erorilor:

$e_i = y_i - b_0 - b_1 x_i$.

În R, estimațiile componentei reziduale se obțin pe baza comenzii:

```
> res=resid(lm(Y~X))
```

O sinteză statistică pentru erorile estimate se obține astfel:

```
> summary(res)
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-10520,0  -2427,0    -969,7       0,0    3177,0   11790,0
> sd(res)
[1] 4035,68
```

Rezultatele de mai sus permit aplicarea testului Student pentru a verifica ipoteza $E(\varepsilon)=0$.

Testul Student

1. Formularea ipotezelor

$H_0: E(\varepsilon)=0$;

$H_1: E(\varepsilon) \neq 0$.

2. Alegerea pragului de semnificație

$\alpha = 0{,}05$.

3. Alegerea testului

Se utilizează statistica Student:

$$t = \frac{\hat{E}(\varepsilon) - E(\varepsilon)}{\sqrt{\hat{V}(\hat{E}(\varepsilon))}}.$$

În condițiile acceptării ipotezei nule, statistica Student este:

$$t = \frac{\hat{E}(\varepsilon)}{\sqrt{\hat{V}(\hat{E}(\varepsilon))}}.$$

4. Valoarea teoretică a testului

Din tabela Student, se citeşte valoarea: $t_{\alpha/2;n-1} = t_{0,025;40} = 1{,}96$.

5. Valoarea calculată a testului

Indicatorii statisticii descriptive pentru erorile estimate sunt prezentaţi în tabelul de mai jos. Pe baza acestora, se calculează:

$$t_{calc} = \frac{E(e_i)}{s_{\hat{E}(\varepsilon)}} = \frac{0}{4035{,}68} = 0.$$

6. Decizia

Comparând valoarea calculată a testului cu valoarea teoretică, rezultă că $t_{calc} \in [-1{,}96;1{,}96]$, ceea ce conduce la decizia de a accepta ipoteza nulă, cu o probabilitate de 0,95. În concluzie, se acceptă ipoteza că media erorilor este zero.

Rezultatul testării se poate obține în R cu ajutorul comenzii:

```
> t.test(res)

One Sample t-test
data: res
t = 3,479e-16,    df = 40,    p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-1273,818;    1273,818
sample estimates:
mean of x
2,192718e-13
```

BCU "CENTRAL UNIVERSITY LIBRARY"

Dacă se compară semnificaţia testului cu pragul de semnificaţie, se observă că *p-value* este egal cu 1 şi este mai mare decât limita *0,05*. Rezultă aceeaşi decizie de a accepta ipoteza că media erorilor nu diferă semnificativ de zero.

7.1.2. Homoscedasticitatea erorilor

Formal, ipoteza se prezintă prin relaţia: $V(\varepsilon_i) = \sigma^2$.

1. Definirea ipotezei

În cazul a două variabile *X*, *Y*, între care există o legătură liniară, regresia este o medie condiţionată definită pe repartiţia bidimensionale (X, Y) şi pe repartiţiile condiţionate de forma: $Y \mid X = x_i$.

Regresia liniară este dată prin relaţia:

$E(Y / X = x_i) = f(x_i) = \beta_0 + \beta_1 x_i$.

La nivelul fiecărei repartiţii condiţionate, se definesc variabilele reziduale:

$\varepsilon_i = y_i - E(Y / X = x_i) = y_i - \beta_0 - \beta_1 x_i$.

Erorile astfel definite sunt *homoscedastice* dacă varianţele acestora sunt egale şi constante.

Exemplu

În figura 7.1, este prezentată repartiţia bidimensională a unui eşantion de familii după consumul şi venitul lunar, exprimate în unităţi monetare. Repartiţiile condiţionate sugerează existenţa *heteroscedasticităţii*.

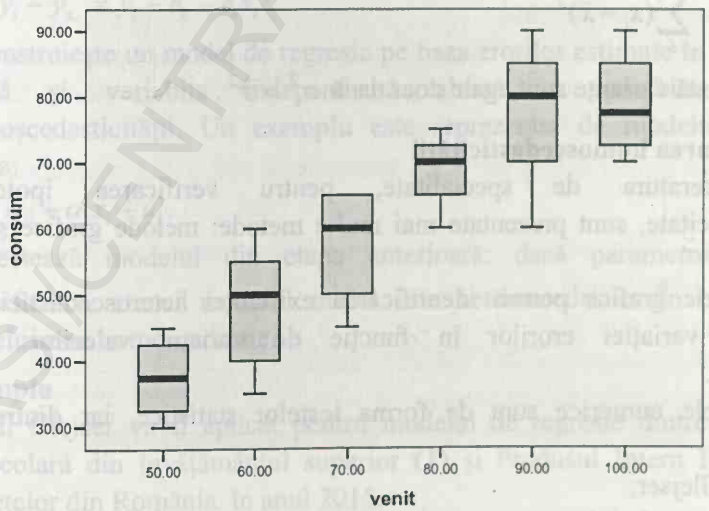


Figura 7.1. *Repartiţia familiilor după venit şi consumul lunar*
Sursa: Realizată de autor (date convenţionale)

BCU IAŞI CENTRAL UNIVERSITY LIBRARY

2. Efectele heteroscedasticităţii

Dacă ipoteza de homoscedasticitate este încălcată, modelul de regresie se numeşte *heteroscedastic*. Efectul încălcării ipotezei de homoscedasticitate este pierderea eficienţei estimatorilor parametrilor modelului de regresie.

Pentru parametrul β_1, de exemplu, se poate arăta că acesta îşi pierde eficienţa, adică estimează parametrul cu o varianţă mai mare, decât în cazul în care ipoteza este verificată.

În acest sens, considerăm relaţia:

$\hat{\beta}_1 = \beta_1 + \sum_i w_i \varepsilon_i$, unde

$$w_i = \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}.$$

În situaţia unui model heteroscedastic, are loc relaţia:

$V(\varepsilon_i) = \sigma_i^2$.

Varianţa estimatorului parametrului β_1 este:

$$V(\hat{\beta}_1) = V(\sum_i w_i \varepsilon_i) = \sum_i w_i^2 V(\varepsilon_i) = \sum_i w_i^2 \sigma_i^2.$$

Aceasta diferă de varianţa estimatorului, în condiţiile în care este respectată ipoteza de homoscedasticitate:

$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}.$$

Cele două varianţe sunt egale doar dacă $\sigma_i^2 = \sigma^2$.

3. Testarea homoscedasticităţii

În literatura de specialitate, pentru verificarea ipotezei de homoscedasticitate, sunt prezentate mai multe metode: metode grafice şi metode numerice.

Metodele grafice permit identificarea existenţei heteroscedasticităţii prin vizualizarea variaţiei erorilor în funcţie de variaţia valorilor variabilei independente.

Metodele numerice sunt de forma testelor statistice, iar dintre acestea menţionăm:

- testul Glejser;
- testul corelaţiei neparametrice între erorile estimate şi variabila independentă;

BCU IASI / CENTRAL UNIVERSITY LIBRARY

- testul Goldfeld-Quandt;
- testul Breusch-Pagan-Godfrey;
- testul White.

a. Testul Glejser

Acest test are la bază un model de regresie între variabila reziduală estimată şi variabila independentă. Forma acestui model indică şi forma heteroscedasticităţii. Ideea de bază a acestui test este că varianţele erorilor σ_i^2 ar putea fi explicate prin valorile variabilei independente.

Observaţii

1. În cazul unui model de regresie multiplă, se identifică acea variabilă independentă, ale cărei valori pot fi asociate cu cele ale varianţei erorilor.

2. Testul Glejser se recomandă doar în cazul în care estimarea modelului de regresie se realizează pe eşantioane mari de date.

Etapele testării

Testarea homoscedasticităţii cu ajutorul testului Glejser presupune parcurgerea următorului demers:

- se construieşte modelul de regresie $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ şi se estimează valorile:
  $y_{x_i} = b_0 + b_1 x_i$;
- pentru modelul propus, se determină erorile estimate:
  $e_i = y_i - y_{x_i} = y_i - b_0 - b_1 x_i$;
- se construieşte un model de regresie pe baza erorilor estimate în valoare absolută şi variabila independentă aleasă ca posibilă sursă a heteroscedasticităţii. Un exemplu este reprezentat de modelul liniar de forma:
  $|e_i| = \alpha_0 + \alpha_1 x_i + u_i$.
- se testează modelul din etapa anterioară: dacă parametrul α_1 este semnificativ, atunci modelul iniţial este heteroscedastic. În caz contrar, modelul este homoscedastic.

Exemplu

Testul Glejser va fi aplicat pentru modelul de regresie dintre variabilele populaţia şcolară din învăţământul superior (Y) şi Produsul Intern Brut (X), la nivelul judeţelor din România, în anul 2015.

Modelul estimat este:

$Y_X = 8273{,}144 + 0{,}528X$.

Variabila reziduală estimată în valoare absolută se obține în R pe baza comenzii:

```
> absres=abs(res)
```

Estimațiile și testele pentru modelul de regresie $|e_i| = \alpha_0 + \alpha_1 x_i + u_i$ se prezintă mai jos:

```
> summary(lm(absres~X))
Call:
lm(formula = absres ~ X)
Coefficients:
             Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)  3,030e+03   7,211e+02    4,202      0,000149 ***
X            1,398e-02   4,968e-02    0,281      0,779832
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 2431 on 39 degrees of freedom
Multiple R-squared: 0,002027,     Adjusted R-squared: -0,02356
F-statistic: 0,07923 on 1 and 39 DF,   p-value: 0,7798
```

Așa cum arată testul Student pentru parametrul α_1, nu avem motive să respingem ipoteza nulă (*p-value =0,77*), ceea ce indică acceptarea ipotezei de homoscedasticitate.

b. Testul corelației neparametrice între erorile estimate și valorile variabilei independente

Acest test este o variantă a testului Glejser și presupune testarea semnificației coeficientului de corelație neparametrică dintre erorile estimate în valoare absolută și variabila independentă.

Pentru un model de regresie liniară simplă, $Y = \beta_0 + \beta_1 X + \varepsilon$, etapele testării sunt următoarele:

- se construiește modelul de regresie, fără a ține seama de ipoteza de homoscedasticitate;
- se estimează erorile pe baza datelor de la nivelul unui eșantion reprezentativ;
- se atribuie ranguri pentru valorile absolute ale erorilor estimate și pentru valorile variabilei independente;
- se calculează coeficientul de corelație neparametric Spearman, pe baza rangurilor valorilor $|e_i|$ și x_i;
- se testează semnificația coeficientului de corelație cu ajutorul testului Student, cu următoarele ipoteze:

BCU BIBLIOTECA UNIVERSITARY

$H_0: \theta = 0$;

$H_1: \theta \neq 0$;

- decizia: modelul este homoscedastic dacă se acceptă ipoteza nulă pentru testul Student din etapa anterioară şi este considerat heteroscedastic dacă se respinge ipoteza nulă.

În acest test, se utilizează:

- estimatorul coeficientului de corelaţie:

$$\hat{\theta} = 1 - 6\frac{\sum_i \hat{d}_i^2}{n(n^2-1)},$$

unde $\hat{d}_i^2$ reprezintă estimatorii diferenţelor dintre ranguri pentru cele două variabile, iar n este volumul eşantionului;

- testul Student:

$$t = \frac{\hat{\theta}\sqrt{n-2}}{\sqrt{1-\hat{\theta}^2}} \sim t(n-2);$$

- valoarea calculată a testului este:

$$t_{calc} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \text{ unde}$$

$r = 1 - 6\frac{\sum_i d_i^2}{n(n^2-1)}$ este estimaţia coeficientului de corelaţie a rangurilor, iar:

$d_i = R_{|e_i|} - R_{x_i}$, adică diferenţa dintre ranguri.

Exemplu

Pentru datele din exemplul anterior, rezultatul testului corelaţiei neparametrice şi comanda în R sunt prezentate mai jos.

```
> cor(absres,X, method="spearman")
```

[1] 0,07979094

Coeficientul de corelaţie Spearman estimat este $r = 0{,}079$.

Testul Student se obţine astfel:

```
> cor.test(absres,X, method="spearman")
```

Spearman's rank correlation rho
data: absres and X
S = 10564, p-value = 0,6189,
alternative hypothesis: true rho is not equal to 0

BCU "CAROL I" CENTRAL UNIVERSITY LIBRARY

sample estimates:
rho
0,07979094

Semnificația testului este *p-value = 0,61* și conduce la decizia de a nu respinge ipoteza nulă. În concluzie, nu se respinge ipoteza de homoscedasticitate pentru modelul de regresie dintre variabilele analizate.

c. Testul Goldfeld-Quandt

Acest test are la bază ideea că între valorile varianței erorilor la nivelul repartițiilor condiționate și valorile variabilei dependente există o legătură pozitivă de forma:

$$\sigma_i^2 = \sigma^2 x_i^2.$$

Pentru realizarea acestui test, se parcurg următoarele etape:

- se ordonează crescător seria de date, la nivelul eșantionului, după variabila X;
- se împarte seria în două părți egale, după omiterea unui set de date din centrul seriei. Sensul omiterii acestor valori este de a obține două subeșantioane de date relativ omogene, cu același volum, pentru valorile mici, respectiv mari, ale variabilei independente.
- se construiesc două modele de regresie pentru cele două seturi de date, utilizând ecuația modelului de regresie inițial;
- se calculează variația reziduală estimată (*RSS*) pentru fiecare model în parte;
- se realizează un test Fisher care compară cele două variații reziduale.

Valoarea calculată a testului este:

$$F_{calc} = \frac{RSS_2}{RSS_1}.$$

Raportul va avea la numărător valoarea *RSS* mai mare, pentru a fi posibilă realizarea testului.

Dacă din seria de date s-au exclus un număr de l date, atunci fiecare subeșantion va fi de volum $(n-l)/2$, iar statistica Fisher va urma o lege de repartiție:

$$F\left(\frac{n-l}{2}-k; \frac{n-l}{2}-k\right).$$

- decizia: dacă testul Fisher este semnificativ statistic, atunci modelul inițial de regresie este heteroscedastic.

Exemplu

Aplicăm testul pe datele oferite de ONU pentru țările din Europa și Asia, în anul 2015, pentru variabilele *Vârsta mediană* (*Y*, ani), *Rata totală de fertilitate* (*X*, copii).

În R, pentru realizarea testului, se utilizează funcția *gqtest* (disponibilă în pachetul *lmtest*), care are sintaxa de mai jos:

```
gqtest(formula, point = 0.5, fraction = 0, alternative = c("greater", "two.sided",
"less"), order.by = NULL, data = list())
```

Formula reprezintă ecuația modelului de regresie. Opțiunea *point* oferă posibilitatea împărțirii seriei inițiale de date după o fracțiune dorită. Implicit, aceasta este 50%. Prin opțiunea *fraction* se specifică numărul de unități din centrul seriei care sunt omise. Se poate introduce și o fracțiune, adică un procent din numărul total. Opțiunea *alternative* vizează ipoteza care se verifică, prima opțiune fiind implicită. Dacă seria nu este ordonată deja, se poate realiza asta prin opțiunea *order.by*.

Comanda și rezultatele testului sunt prezentate mai jos:

```
> gqtest(lm(Y~X), point = 0.5, fraction = 0, order.by =X )

Goldfeld-Quandt test
data: lm(Y ~ X)
GQ = 1,2817,  df1 = 49,  df2 = 49,  p-value = 0,194
alternative hypothesis: variance increases from segment 1 to 2
```

S-au construit două regresii pentru două subeșantioane de câte 49 de unități.

Valoarea calculată a testului este $GQ = 1{,}28$, iar semnificația testului este *p-value* = 0,19.

Decizia: deoarece semnificația testului este mai mare decât riscul asumat de 0,05, nu se respinge ipoteza de homoscedasticitate.

Observație

Testele utilizate mai sus se pot aplica în special pentru modelele de regresie simplă. Pentru cazul modelelor multiple, propunem câteva teste în cele ce urmează.

d. Breusch-Pagan-Godfrey

Ideea de bază a acestui test este aceea de a construi un model de regresie care are ca variabilă dependentă variabila eroare estimată la pătrat, iar ca variabile independente pot fi utilizate variabilele din modelul de regresie de bază pentru care se testează heteroscedasticitatea.

Astfel, pentru un model de forma $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, testarea homoscedasticității presupune următorul demers:

- se construiește modelul de regresie de bază și se estimează parametrii acestuia;
- se estimează erorile pe baza datelor de la nivelul unui eșantion reprezentativ;
- se construiește modelul auxiliar de regresie:

$e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i$;

- pe baza estimației coeficientului de determinație din acest model (R_a^2), se construiește un test Chi-pătrat cu $k-1$ grade de libertate, unde k este numărul de parametri din modelul auxiliar;
- valoarea calculată a testului este de forma:

$\chi^2_{calc} = n \cdot R_a^2$;

- decizia de a accepta ipoteza de homoscedasticitate se ia pe baza comparării valorii teoretice din tabele ($\chi^2_{\alpha,k-1}$) cu cea calculată în etapa anterioară.

Testarea în R se realizează cu ajutorul funcției *bptest*, disponibilă în pachetul *lmtest*. Sintaxa funcției este următoarea:

```
> bptest(formula, varformula = NULL, studentize = TRUE, data = list())
```

Opțiunea *varformula* dă posibilitatea construirii unui test care ia în considerare un număr mai redus de variabile independente. Următoarea opțiune (*studentize*) dă posibilitatea construirii unei alte forme studentizate a testului (varianta Koenker).

Exemplu

Considerăm datele disponibile la nivelul județelor României, în 2015, pentru variabilele *Rata brută de natalitate* (*Y*, născuți vii la 1000 de locuitori), *Rata brută de nupțialitate* (X_1, căsătorii la 1000 de locuitori) și *Rata brută de divorțialitate* (X_2, divorțuri la 1000 de locuitori).

Pentru modelul de regresie dintre aceste variabile, testul de homoscedasticitate este realizat cu ajutorul programului R. Rezultatele modelării și ale testului sunt prezentate mai jos.

```
> summary(lm(Nat~Cas+Div))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6,6019	0,9170	7,199	1,14e-08 ***
Cas	0,6224	0,1621	3,840	0,000441 ***
Div	-0,8646	0,4498	-1,922	0,061910 .

```
> bptest(lm(Nat~Cas+Div))
```

studentized Breusch-Pagan test

data: lm(Nat ~ Cas + Div)
BP = 4,2446, df = 2, p-value = 0,1198

Rezultatele de mai sus permit luarea deciziei de a nu respinge ipoteza de homoscedasticitate. Se observă că semnificaţia testului este mai mare decât pragul de 0,05.

Observaţie

Putem realiza în R testul Breusch-Pagan-Godfrey pas cu pas, urmând etapele:

- se estimează modelul de regresie dintre variabile;
- se estimează erorile de modelare;
- se construieşte variabila reziduală la pătrat;
- se construieşte modelul de regresie auxiliar între pătratul reziduurilor şi predictorii din modelul iniţial şi se estimează coeficientul de determinaţie auxiliar;
- se calculează valoarea testului $\chi^2_{calc} = n \cdot R^2_a$;
- se ia decizia comparând valoarea calculată a testului cu valoarea teoretică $\chi^2_{\alpha,k-1}$.

Comenzile şi rezultatele sunt:

```
> lm(Nat~Cas+Div)
> res=resid(lm(Nat~Cas+Div))
> res2=res*res
> summary(lm(res2~Cas+Div))
```

Multiple R-squared: 0,1011, Adjusted R-squared: 0,05496
F-statistic: 2,192 on 2 and 39 DF, p-value: 0,1252

$\chi^2_{calc} = 42 \cdot 0{,}10 = 4{,}2$

$\chi^2_{\alpha,k-1} = \chi^2_{0{,}05;2} = 5{,}9$.

Concluzia este aceeaşi, adică nu se respinge ipoteza de homoscedasticitate.

e. Testul White

Acest test se aseamănă cu cel precedent prin modul de concepţie, însă regresia auxiliară are o altă formă.

Pentru un model de regresie $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, demersul testării presupune aceiaşi paşi, iar testul se construieşte cu ajutorul următoarei ecuaţii de regresie auxiliară:

$$e^2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2 + \alpha_4 X_1^2 + \alpha_5 X_2^2 + u\,.$$

BCU CENTRAL UNIVERSITY LIBRARY

Testarea se realizează în mod analog cazului anterior, cu ajutorul unui test Chi-pătrat de k - 1 grade de libertate, unde k este numărul de parametri din model.

Valoarea calculată $\chi^2_{calc} = n \cdot R_a^2$ se obține pe baza modelului de mai sus și se compară cu valoarea teoretică din tabel, pentru un prag de semnificație fixat.

În R, se poate aplica aceeași funcție *bptest*, iar la *formula* se introduce expresia din partea dreaptă a regresiei auxiliare.

Exemplu

Considerăm datele de la exemplul precedent. Rezultatele sunt:

```
> bptest(lm(Nat~Cas+Div+Cas*Div+I(Cas^2)+I(Div^2)))

studentized Breusch-Pagan test
data: lm(Nat ~ Cas + Div + Cas * Div + I(Cas^2) + I(Div^2))
BP = 8,7644,   df = 5,   p-value = 0,1188
```

Rezultatul indică o semnificație a testului cu o valoare mai mare de 0,05. Concluzia este că nu putem respinge ipoteza de homoscedasticitate pentru modelul de regresie analizat.

4. Corectarea heteroscedasticității

Dacă în urma testării ipotezei de homoscedasticitate s-a constatat că ipoteza nu se verifică, se impune corectarea modelului.

Acest lucru este posibil în funcție de următoarele două situații: parametrii σ_i^2 sunt cunoscuți și parametrii σ_i^2 nu sunt cunoscuți.

a. σ_i^2 sunt cunoscuți

Corecția heteroscedasticității este aplicată modelului de regresie liniară simplă:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i .$$

În condițiile în care se cunosc parametrii σ_i^2, modelul poate fi transformat prin relația:

$$\frac{y_i}{\sigma_i} = \frac{\beta_0}{\sigma_i} + \beta_1 \frac{x_i}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}.$$

Noul model de regresie este:

$$y_i^* = \beta_0^* \frac{1}{\sigma_i} + \beta_1^* x_i^* + \varepsilon_i^*, \text{ unde:}$$

$$y_i^* = \frac{y_i}{\sigma_i},$$

BCU IASI CENTRAL UNIVERSITY LIBRARY

$$x_i^* = \frac{x_i}{\sigma_i},$$

$$\varepsilon_i^* = \frac{\varepsilon_i}{\sigma_i}.$$

Se poate demonstra că acest model este homoscedastic, deoarece varianţa erorilor este aceeaşi pentru fiecare repartiţie condiţionată şi este constantă:

$$V(\varepsilon_i^*) = V(\frac{\varepsilon_i}{\sigma_i}) = \frac{1}{\sigma_i^2} V(\varepsilon_i) = 1.$$

Observaţie

Corectarea heteroscedasticităţii presupune ponderarea modelului iniţial cu variabila $\frac{1}{\sigma_i}$. Estimarea parametrilor pentru modelul corectat se poate realiza prin aplicarea metodei celor mai mici pătrate, care, în acest caz, poartă denumirea de *metoda celor mai mici pătrate ponderate* (*method of weighted least squares*).

b. σ_i^2 sunt necunoscuţi

Dacă nu sunt cunoscuţi parametrii σ_i^2, corectarea modelului se poate realiza prin utilizarea unor transformări care au la bază diferite ipoteze funcţionale între parametrii σ_i^2 şi variabila independentă. Asemenea relaţii pot fi detectate cu ajutorul testului Glejser. Analizăm două situaţii.

b1. Un prim exemplu des întâlnit este corecţia modelului pe baza relaţiei:

$$\sigma_i^2 = \sigma^2 x_i^2.$$

În acest caz, modelul corectat are forma:

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\varepsilon_i}{x_i}.$$

Prin transformare, se obţine modelul:

$$y_i^* = \beta_0^* \frac{1}{x_i} + \beta_1^* + \varepsilon_i^*, \text{ în care: } V(\varepsilon_i^*) = \sigma^2.$$

Această metodă utilizează ca variabilă de ponderare a modelului iniţial variabila $\frac{1}{X}$.

Observaţie

În R, se poate construi un model cu ajutorul *metodei celor mai mici pătrate ponderate* prin utilizarea opţiunii *weights* în comanda *lm*.

b2. Prezentăm un alt exemplu de corecție a heteroscedasticității ce se poate aplica și modelelor de regresie multiplă. Această metodă presupune corecția modelului pe baza relației: $\sigma_i^2 = \sigma^2 h(x_i)$, unde funcția h poate fi o exponențială a predictorilor din modelul de regresie inițial. Să considerăm cazul unui model multiplu cu două variabile independente.

Pentru erorile heteroscedastice, se poate scrie relația:

$\ln \varepsilon_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i$, unde u_i respectă ipotezele clasice.

Pentru corectarea heteroscedasticității cu ajutorul *metodei celor mai mici pătrate ponderate,* se parcurg etapele:

- se construiește modelul de regresie de bază și se estimează parametrii acestuia;
- se estimează erorile modelului, pentru care se verifică ipoteza de homoscedasticitate;
- în cazul în care erorile sunt heteroscedastice, se construiește modelul auxiliar de regresie:

  $g_i = \ln e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i$ și se estimează valorile g_i;
- variabila de ponderare este: $h_i = \exp(g_i)$;
- modelul inițial este corectat cu ponderea $\frac{1}{h_i}$, iar noul model obținut este unul homoscedastic: $\frac{y_i}{h_i} = \frac{\beta_0}{h_i} + \beta_1 \frac{x_{1i}}{h_i} + \beta_2 \frac{x_{2i}}{h_i} + \frac{\varepsilon_i}{h_i}$.

7.1.3. Normalitatea erorilor

Formal, această ipoteză este prezentată prin relația: $\varepsilon_i \sim N(0, \sigma^2)$.

1. Definirea ipotezei

Estimatorii parametrilor modelului de regresie sunt combinații liniare în care apare variabila eroare. Dacă este respectată ipoteza de normalitate a erorilor, estimatorii parametrilor modelului de regresie urmează, de asemenea, o lege de repartiție normală.

Pentru modelul de regresie liniară simplă, $Y = \beta_0 + \beta_1 X + \varepsilon$, prin metoda celor mai mici pătrate, se obțin estimatorii:

$$\hat{\beta}_1 = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} = \sum_i \frac{nx_i - \sum_i x_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} \cdot y_i,$$

BCU ASCENTRALUNIVERSITARA LIBRARY

$\hat{\beta}_0 = \hat{\bar{y}} - \hat{\beta}_1 \bar{x} = \sum_i y_i (\frac{1}{n} - \bar{x} w_i)$, unde

$w_i = \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$, iar

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Dacă $\varepsilon_i \sim N(0, \sigma^2)$, au loc relaţiile:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right) \text{ sau } \hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right),$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right)\right) \text{ sau } \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right), \text{ unde}$$

σ^2 este dispersia erorilor.

2. Efectele încălcării ipotezei

Dacă erorile de modelare nu urmează o lege de repartiţie normală, atunci estimatorii construiţi pe baza metodei celor mai mici pătrate nu urmează o lege de repartiţie normală.

Pentru eşantioane de volum mare, proprietatea de normalitate este atinsă asimptotic.

3. Verificarea ipotezei de normalitate a erorilor

Testarea normalităţii variabilei reziduale se poate realiza în două moduri: pe cale grafică şi cu ajutorul testelor numerice.

a. Evaluarea pe cale grafică a normalităţii erorilor

Abordarea grafică a repartiţiei variabilei eroare corespunzătoare unui model de regresie se poate realiza cu ajutorul: histogramei erorilor, diagramei QQ-Plot, diagramei Box-Plot etc.

Histograma erorilor

Cu ajutorul histogramei erorilor, se poate realiza o comparaţie vizuală a formei repartiţiei erorilor cu forma repartiţiei normale. Existenţa unor abateri sau a unor diferenţe semnificative între cele două repartiţii conduce la presupunerea încălcării ipotezei de normalitate a erorilor.

BCU CENTRAL UNIVERSITY LIBRARY

Exemplu

Pentru datele privind populația școlară din învățământul superior (Y) și Produsul Intern Brut (X), la nivelul județelor României, în anul 2015, s-a construit un model de regresie liniară simplă. În R, variabila reziduală estimată se obține cu ajutorul comenzii:

```
> res=resid(lm(Y~X))
```

Pentru a reprezenta grafic histograma erorilor și curba normală, se utilizează comenzile:

```
> hist(res, freq=F)
> curve(dnorm(x, mean(res), sd(res)), add=T)
```

Așa cum arată și figura 7.2, nu există o abatere semnificativă a repartiției erorilor de la forma repartiție normale.

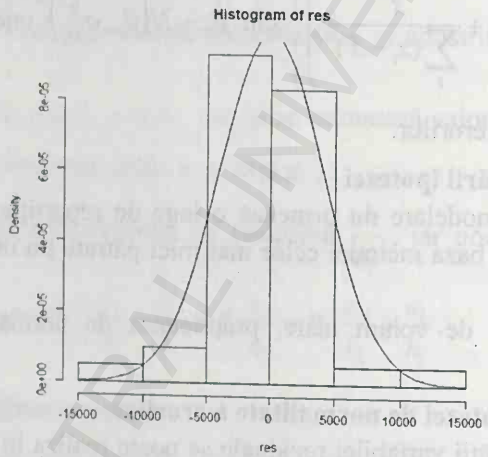


Figura 7.2. *Repartiția erorilor estimate*

Sursa: Realizată de autor pe baza datelor oferite de INS, 2016

Diagrama QQ-Plot

Această diagramă compară vizual frecvențele relative cumulate observate ale erorilor cu cele așteptate sau cu probabilitățile corespunzătoare obținute cu ajutorul tabelei lui Laplace sau a repartiției normale.

Frecvențele cumulate teoretice sau așteptate se distribuie sub forma unei drepte (dreapta lui Henry) – este diagonala din pătratul probabilităților, pătrat în care valoarea minimă este zero și cea maximă este unu. Frecvențele cumulate observate se pot afla foarte aproape de această dreaptă și de-a lungul acesteia sau se pot abate în mod semnificativ de o parte și de alta a dreptei frecvențelor așteptate.

BCU "MIHAIL EMINESCU" CENTRAL UNIVERSITY LIBRARY

În primul caz, putem considera că repartiția erorilor respectă condiția de normalitate, iar în al doilea, această ipoteză este încălcată.

Exemplu

Pentru același model prezentat mai sus, s-a obținut diagrama QQ-Plot.

În R, sunt necesare comenzile:

```
> qqnorm(res)
> qqline(res)
```

Așa cum se observă din diagrama 7.3, între cele două repartiții există mici diferențe, adică erorile modelului de regresie dintre cele două variabile analizate par să se abată de la forma repartiției normale. Evident, o decizie clară cu privire la ipoteza de normalitate presupune realizarea unui test.

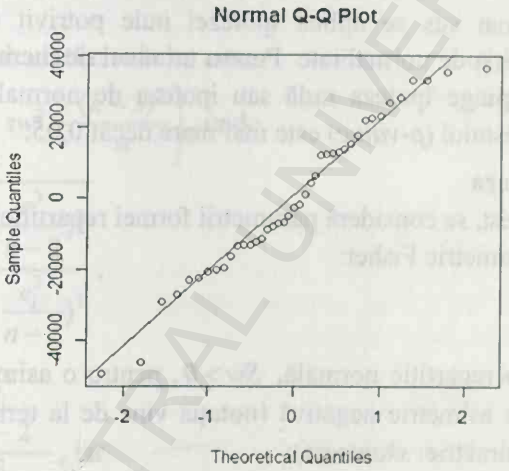


Figura 7.3. *Diagrama QQ-Plot a erorilor*
Sursa: Realizată de autor pe baza datelor oferite de INS, 2016

b. Teste numerice

Testarea normalității repartiției erorilor se poate realiza cu un test neparametric clasic, cum ar fi testul Chi-pătrat sau testul Kolmogorov - Smirnov. Pe lângă acestea, în literatura de specialitate, se întâlnește un test care se construiește pe baza parametrilor formei unei repartiții: asimetria și boltirea. Acesta este *testul Jarque-Bera*, după numele statisticienilor care l-au elaborat.

Testul Kolmogorov - Smirnov

Statistica sau testul Kolmogorov - Smirnov se bazează pe maximul diferenței, în valoare absolută, dintre două tipuri de frecvențe: empirice și teoretice.

BCU "CAROL I" CENTRAL UNIVERSITY LIBRARY

Frecvențele empirice se obțin pe baza estimațiilor erorilor asociate modelului de regresie.

Exemplu

Pentru erorile estimate în cazul modelului de regresie prezentat la exemplul anterior, se aplică testul Kolmogorov - Smirnov. În R, comanda și rezultatele sunt redate mai jos (se compară distribuția erorilor cu o distribuție normală de medie zero și abatere standard egală cu estimație la nivel de eșantion).

```
> ks.test(res, pnorm, 0, 4000)

One-sample Kolmogorov-Smirnov test
data: res
D = 0,118        p-value = 0,56
alternative hypothesis: two-sided
```

Rezultatele de mai sus se aplică ipotezei nule potrivit căreia variabila reziduală respectă condiția de normalitate. Pentru un nivel de încredere de 0,95, se ia decizia de a nu respinge ipoteza nulă sau ipoteza de normalitate a erorilor, deoarece semnificația testului (*p-value*) este mai mare decât 0,05.

Testul Jarque-Bera

În cazul acestui test, se consideră parametrii formei repartiției erorilor:

- coeficientul de asimetrie Fisher:

$Sw = \frac{\mu_3}{\sigma^3}$, unde

$Sw = 0$, pentru o repartiție normală, $Sw > 0$, pentru o asimetrie pozitivă și $Sw < 0$, pentru o asimetrie negativă (notația vine de la termenul din limba engleză pentru asimetrie: *skewness*);

- coeficientul de boltire:

$K = \frac{\mu_4}{\mu_2^2}$, unde

$K = 3$, pentru o repartiție normală, $K < 3$, pentru o repartiție aplatizată și $K > 3$, pentru o repartiție cu boltire (notația vine de la termenul din limba engleză pentru boltire: *kurtosis*).

Estimatorii pentru cei doi parametri sunt:

$$\hat{S}w = \sqrt{\frac{(\sum_i \frac{\hat{\varepsilon}_i^3}{n-2})^2}{(\sum_i \frac{\hat{\varepsilon}_i^2}{n-2})^3}}$$, respectiv

$$\hat{K} = \frac{\sum_i \frac{\hat{\varepsilon}_i^4}{n-2}}{(\sum_i \frac{\hat{\varepsilon}_i^2}{n-2})^2}.$$

Statistica Jarque-Bera are relaţia:

$$JB = \frac{n}{6}\left(\hat{Sw}^2 + \frac{(\hat{K}-3)^2}{4} \right) \sim \chi^2(2),$$

adică urmează o lege de repartiţie Chi-pătrat de două grade de libertate.

Testarea se realizează cu ajutorul datelor disponibile, pe baza cărora se calculează erorile estimate. La nivelul acestei repartiţii, se obţin estimaţiile pentru parametrii formei repartiţiei erorilor şi se calculează o valoare a testului Jarque-Bera.

Valoarea calculată a testului este:

$$JB_{calc} = \frac{n}{6}\left(sw^2 + \frac{(k-3)^2}{4} \right), \text{ unde}$$

$$sw = \sqrt{\frac{(\sum_i \frac{e_i^3}{n-2})^2}{(\sum_i \frac{e_i^2}{n-2})^3}},$$

$$k = \frac{\sum_i \frac{e_i^4}{n-2}}{(\sum_i \frac{e_i^2}{n-2})^2}, \text{ iar}$$

$$e_i = y_i - b_0 - b_1 x_i.$$

Ipoteza de normalitate a erorilor se admite în cazul în care valoarea calculată a testului este mai mică decât valoarea teoretică pentru o distribuţie Chi-pătrat de două grade de libertate şi un prag de semnificaţie α specificat, adică $JB_{calc} \le \chi^2_{\alpha,2}$.

Dacă $JB_{calc} > \chi^2_{\alpha,2}$, se respinge ipoteza nulă, adică ipoteza de normalitate a erorilor, cu o probabilitate egală cu $1-\alpha$.

Exemplu

Utilizăm modelul de regresie prezentat la exemplul anterior.

Putem obţine estimaţiile celor doi parametri ai formei cu ajutorul comenzilor (se utilizează pachetul *moments*):

BCU "ASSENTRAL UNIVERSITY LIBRARY

```
> skewness(res)
[1] 0,2548646
> kurtosis(res)
[1] 3,807929
```

Estimațiile parametrilor formei repartiției erorilor sunt:
$sw = 0,2548$ și $k = 3,8079$.

Valoarea statisticii Jarque-Bera este următoarea:

$$JB_{calc} = \frac{n}{6}\left(sw^2 + \frac{(k-3)^2}{4}\right) = \frac{41}{6}(0,064 + 0,162) = 1,55.$$

Potrivit tabelei Chi-pătrat, valoarea teoretică a testului este: $\chi^2_{0,05;2} = 5,99$.

În concluzie, $JB_{calc} < \chi^2_{\alpha,2}$, ceea ce conduce la decizia de a nu respinge ipoteza de normalitate a erorilor.

În R, se poate construi automat testul, utilizând comanda:

```
> jarque.bera.test(res)

Jarque Bera Test
data: res
X-squared = 1,559,  df = 2,  p-value = 0,4586
```

Deoarece semnificația testului este mai mare decât 0,05, se ia decizia de a nu respinge ipoteza de normalitate a distribuției erorilor.

7.1.4. Necorelarea erorilor

Această ipoteză se prezintă formal prin relația: $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$. Cele patru dimensiuni ale analizei fiecărei ipoteze sunt prezentate mai jos.

Această ipoteză este importantă în special în cazul seriilor de timp și al modelelor construite pentru serii spațiale, situație în care trecerea de la o unitate de timp sau de spațiu la alta poate însemna și transmiterea unui efect sau influențe la nivelul valorilor variabilei reziduale (pentru seriile spațiale vorbim despre un *efect de vecinătate*). În acest subcapitol, tratăm subiectul fără a face trimitere direct la un tip de serie statistică, accentul fiind pus pe elementele de ordin metodologic.

1. Definirea ipotezei

Variabilele aleatoare reziduale definite la nivelul repartițiilor condiționate (de forma $Y|X = x_i$) pot fi independente sau corelate.

Ipoteza de necorelare a erorilor se referă la lipsa unei corelații între variabilele reziduale sau la faptul că eroarea asociată unei valori a variabilei

dependente nu este influenţată de eroarea asociată altei valori a variabilei dependente.

În cazul în care ipoteza de independenţă nu se verifică, modelul de regresie înregistrează o autocorelare a erorilor sau o corelaţie serială. Formal, încălcarea ipotezei înseamnă existenţa unei relaţii de forma:

$$\text{cov}(\varepsilon_i, \varepsilon_j) \neq 0 .$$

Dacă se admite ipoteza că media erorilor este zero, relaţia anterioară este echivalentă cu $E(\varepsilon_i \cdot \varepsilon_j) \neq 0$.

Autocorelarea erorilor poate fi cauzată de:

- lipsa unei specificaţii adecvate a formei modelului de regresie;
- lipsa din model a uneia sau a mai multor variabile semnificative;
- sistematizarea şi pregătirea datelor pentru prelucrare;
- inerţia fenomenelor în timp şi decalajul, în cazul seriilor de timp;
- efectul de vecinătate, în cazul seriilor spaţiale.

În condiţiile încălcării ipotezei de necorelare a erorilor, se poate considera că între erori există o relaţie de forma:

$$\varepsilon_i = \rho \varepsilon_{i-1} + u_i ,$$

unde u_i reprezintă o variabilă pur aleatoare (numită „zgomot alb") care respectă ipotezele modelului clasic de regresie.

Parametrul ρ este coeficientul de autocorelaţie între ε_i şi ε_{i-1} şi este definit de relaţia:

$$\rho = \frac{\text{cov}(\varepsilon_i, \varepsilon_{i-1})}{\sigma_i \sigma_{i-1}} .$$

Pentru variabila u_i au loc relaţiile:

- $E(u_i) = 0$, media erorilor este nulă;
- $V(u_i) = \sigma_u^2$, homoscedasticitatea erorilor;
- $u_i \sim N(0, \sigma_u^2)$, normalitatea erorilor;
- $\text{cov}(u_i, u_j) = 0$, necorelarea erorilor.

Dacă există autocorelare a erorilor pentru modelul de regresie, iar celelalte ipoteze se respectă, intensitatea legăturii dintre erori este măsurată prin:

$$\rho = \frac{\text{cov}(\varepsilon_i, \varepsilon_{i-1})}{\sigma^2} = \frac{\sum_i \varepsilon_i \cdot \varepsilon_{i-1}}{\sum_i \varepsilon_i^2} .$$

În aceste condiţii, au loc relaţiile:

$$\sigma^2 = V(\varepsilon_i) = V(\rho\varepsilon_{i-1} + u_i) = \rho^2\sigma^2 + \sigma_u^2,$$

$$\sigma^2 = \frac{\sigma_u^2}{1-\rho^2}.$$

Din ultima relație, se poate interpreta semnificația coeficientului de corelație dintre erori:

- $\rho = 0$, erorile nu sunt corelate, ci se comportă ca variabile pur aleatoare;
- $\rho \neq 0$, între erori există o legătură, iar dispersia acestora este mai mare decât dispersia unei variabile pur aleatoare;
- $\rho = 1$, erorile sunt perfect corelate, iar modelarea nu se poate realiza.

Observație

Măsurarea intensității corelației dintre erori se poate realiza și pentru un decalaj cu ordin mai mare decât unu. Pentru astfel de situații, se definește funcția de autocorelație de ordin *k*, potrivit relației:

$$f(k) = \frac{\text{cov}(\varepsilon_i, \varepsilon_{i-k})}{\sigma_i \sigma_{i-k}} = \frac{\text{cov}(\varepsilon_i, \varepsilon_{i-k})}{\sigma^2}.$$

2. Efectele încălcării ipotezei

În condițiile existenței autocorelării erorilor, este afectată calitatea estimațiilor obținute prin metoda celor mai mici pătrate.

Se poate demonstra că, prin aplicarea metodei celor mai mici pătrate, pentru parametrul β_0, se obține un estimator neeficient.

Considerăm modelul de regresie liniară simplă:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Pentru erori, considerăm că are loc relația:

$\varepsilon_i = \rho\varepsilon_{i-1} + u_i$.

Din acest model, se poate observa că eroarea pur aleatoare se obține ca o diferență de tipul:

$\varepsilon_i - \rho\varepsilon_{i-1} = u_i$.

Aceasta este o *quasi-diferență*, care se obține cu ajutorul coeficientului de corelație de ordinul întâi dintre erorile modelului.

Pe baza acestei quasi-diferențe, se poate construi un model de regresie transformat, în care variabila aleatoare să fie tocmai variabila u_i.

În acest sens, se construiește modelul de regresie cu un decalaj:

$y_{i-1} = \beta_0 + \beta_1 x_{i-1} + \varepsilon_{i-1}$.

Dacă acest ultim model se înmulțește cu ρ și se scade din modelul inițial, rezultă modelul:

BCU "LUCIAN BLAGA" UNIVERSITY LIBRARY

$$y_i - \rho \cdot y_{i-1} = \beta_0(1-\rho) + \beta_1(x_i - \rho \cdot x_{i-1}) + u_i .$$

Acest ultim model admite ca variabilă reziduală o variabilă aleatoare pură, adică admite ipoteza de necorelare a erorilor.

Modelul de mai sus se numeşte model de *quasi-diferenţă* şi se poate scrie astfel:

$y_i^* = \beta_0^* + \beta_1^* x_i^* + u_i$, unde

$\beta_0^* = \beta_0(1-\rho)$;

$\beta_1^* = \beta_1$;

$x_i^* = x_i - \rho \cdot x_{i-1}$;

$y_i^* = y_i - \rho \cdot y_{i-1}$.

Acest model respectă ipotezele modelului clasic de regresie, iar prin aplicarea metodei celor mai mici pătrate, ne oferă un alt estimator pentru parametrul β_0, care este nedeplasat şi eficient.

3. Testarea autocorelării erorilor

Autocorelarea erorilor se poate testa cu ajutorul mai multor teste, dintre care vom prezenta următoarele două teste mai des utilizate: *testul Durbin - Watson* şi *testul Runs*.

a. Testul Durbin - Watson

Acesta presupune testarea semnificaţiei coeficientului de corelaţie de ordinul întâi dintre erori. În cazul în care coeficientul este semnificativ statistic, modelul de regresie admite fenomenul de autocorelare a erorilor, iar în caz contrar, ipoteza de necorelare este respectată.

Testul Durbin - Watson se realizează prin parcurgerea etapelor prezentate mai jos.

1. Formularea ipotezelor

$H_0: \rho = 0$ (erorile nu sunt autocorelate);

$H_1: \rho \neq 0$ (există autocorelare a erorilor).

2. Alegerea pragului de semnificaţie (de regulă, se consideră $\alpha = 0{,}05$).

3. Alegerea testului

Se utilizează statistica Durbin-Watson:

$$DW = d = \frac{\sum_i (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_i \hat{\varepsilon}_i^2} .$$

Dacă se presupune existența autocorelației de forma:

$\varepsilon_i = \rho\varepsilon_{i-1} + u_i$,

statistica *DW* se mai poate scrie astfel:

$$d = \frac{\sum_i \hat{\varepsilon}_i^2 - 2\sum_i \hat{\varepsilon}_i\hat{\varepsilon}_{i-1} + \sum_i \hat{\varepsilon}_{i-1}^2}{\sum_i \hat{\varepsilon}_i^2} \cong 2\frac{\sum_i \hat{\varepsilon}_i^2 - \sum_i \hat{\varepsilon}_i\hat{\varepsilon}_{i-1}}{\sum_i \hat{\varepsilon}_i^2} =$$

$$= 2\left(1 - \frac{\sum_i \hat{\varepsilon}_i\hat{\varepsilon}_{i-1}}{\sum_i \hat{\varepsilon}_i^2}\right) = 2(1-\hat{\rho}).$$

Estimatorul coeficientului de corelație a erorilor este:

$$\hat{\rho} = \frac{\sum_i \hat{\varepsilon}_i\hat{\varepsilon}_{i-1}}{\sum_i \hat{\varepsilon}_i^2}$$

și respectă condiția: $-1 \le \hat{\rho} \le 1$.

Dacă $d = 2(1-\hat{\rho})$, valorile statisticii *DW* sunt cuprinse în intervalul: $0 \le d \le 4$.

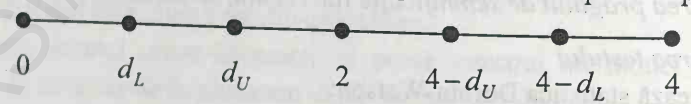
Interpretare

- $\hat{\rho} = -1 \Rightarrow d = 4$, între erori există autocorelare negativă maximă;
- $\hat{\rho} = 1 \Rightarrow d = 0$, între erori există autocorelare pozitivă maximă;
- $\hat{\rho} = 0 \Rightarrow d = 2$, nu există autocorelare a erorilor.

4. Determinarea valorii teoretice a testului

Din tabela Durbin-Watson, se citesc valorile critice ale statisticii *DW*, în funcție de pragul de semnificație și de volumul eșantionului.

În tabele, sunt prezentate două valori critice, notate cu d_L (limita inferioară) și d_U (limita superioară), pentru diverse valori ale pragului de semnificație și ale volumului eșantionului. În funcție de aceste valori critice, se determină următoarele intervale, care permit luarea deciziei de respingere sau de acceptare a ipotezei nule:



- $(0; d_L)$ este o regiune de respingere, erorile înregistrează o autocorelare pozitivă;

BCU IASI - THE UNIVERSITY LIBRARY

- $(d_L; d_U)$ şi $(4-d_U; 4-d_L)$ sunt regiuni de nedeterminare şi nu permit luarea unei decizii asupra existenţei autocorelării erorilor;
- $(d_U; 4-d_U)$ este o regiune de acceptare a ipotezei nule, erorile nu sunt autocorelate;
- $(4-d_L; 4)$ este o regiune de respingere, erorile înregistrează o autocorelare negativă.

5. Determinarea valorii calculate a testului

Pe baza datelor de la nivelul unui eşantion, se calculează o valoare a statisticii Durbin-Watson:

$$d_{calc} = \frac{\sum_i (e_i - e_{i-1})^2}{\sum_i e_i^2}.$$

6. Decizia

Decizia de a accepta sau a respinge ipoteza nulă se ia în urma comparării valorii calculate a testului cu valorile critice din tabela Durbin-Watson, adică în funcţie de poziţia valorii calculate în una dintre regiunile specificate la punctul 4.

Observaţie

Testul Durbin - Watson nu realizează decât un test asupra existenţei unei autocorelări de ordinul întâi între termenii variabilei eroare. Pentru a lua în considerare posibilele corelaţii între termenii cu un decalaj de ordin mai mare decât unu, se poate considera un model de forma:

$$\varepsilon_i = \rho' \varepsilon_{i-1} + \rho'' \varepsilon_{i-2} + ... + \rho^{(p)} \varepsilon_{i-p} + u_i.$$

Decizia asupra încălcării ipotezei de necorelare a erorilor se ia în urma testării valorilor funcţiei de autocorelaţie pentru decalaje de diverse ranguri.

Exemplu

Pentru datele privind populaţia şcolară din învăţământul superior (Y) şi Produsul Intern Brut (X), la nivelul judeţelor României, în anul 2015, s-a construit un model de regresie liniară simplă.

După estimarea variabilei reziduale, se aplică testul Durbin - Watson cu ajutorul comenzii de mai jos (pachetul *cars*):

```
> dwt(res)
[1] 2,119062
```

Aşadar, valoarea calculată a statisticii Durbin - Watson este: $d_{calc} = 2,11$.

Din tabela Durbin - Watson, pentru un prag de semnificație de 0,05, pentru un model de regresie cu doi parametri și un eșantion de volum $n=41$, se citesc cele două valori critice:

$d_L = 1,44$;

$d_U = 1,54$.

În concluzie, d_{calc} aparține intervalului $(d_U; 4-d_U)$, ceea ce conduce la decizia de a nu respinge ipoteza nulă, adică se consideră că între erorile asociate modelului de regresie nu există autocorelare.

b. Testul Runs

Valorile variabilei aleatoare eroare pot fi privite ca seturi de valori care se succed în funcție de semnul lor. Succesiunea acestor secvențe de date poate fi aleatoare sau poate avea o anumită regularitate sau ordine. Un *run* este o astfel de secvență de valori de același semn ale variabilei eroare.

În cazul independenței erorilor, succesiunea de *runs* este aleatoare, iar numărul acestora este distribuit normal. În caz contrar, numărul de *runs* nu este distribuit normal, iar secvențele apar într-o anumită ordine.

Pentru testare, se utilizează următoarele notații:

n – volumul eșantionului;

e_i – valorile estimate ale erorilor;

n_1 – numărul de valori pozitive ale valorilor e_i;

n_2 – numărul de valori negative ale e_i;

$n_1 + n_2 = n$;

K – numărul de *runs*, variabilă aleatoare care are următorii parametri:

$$E(K) = 2\frac{n_1 n_2}{n_1 + n_2} + 1,$$

$$V(K) = \sigma_k^2 = 2n_1 n_2 \frac{2n_1 n_2 - n_1 - n_2}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}.$$

Etapele testării

Pentru realizarea testului, se parcurg etapele:

1. Formularea ipotezelor

H_0: K este distribuit normal (nu există autocorelare a erorilor);

H_1: K nu este distribuit normal (ipoteza este încălcată).

2. Alegerea pragului de semnificație (de obicei, este *0,05*).

3. Alegerea testului

BCU ACADEMIA UNIVERSITARĂ LIBRARY

Pentru testare, se utilizează o statistică Student:

$$t = \frac{K - E(K)}{\hat{\sigma}_K}.$$

4. Valoarea teoretică a testului

Pentru un prag de semnificaţie de 5%, se citeşte din tabel o valoare teoretică a testului Student $t(n-2)$.

5. Valoarea calculată a testului

Valoarea calculată se obţine pe baza estimaţiilor mediei şi varianţei variabilei K.

6. Decizia de a accepta ipoteza nulă se ia în cazul în care valoarea calculată se află în intervalul: $[-t_{0,025;n-2}; t_{0,025;n-2}]$.

Exemplu

Utilizăm modelul de regresie prezentat la exemplul anterior.

În R, testul se realizează cu ajutorul comenzii (în pachetul *randtests*):

```
> runs.test(res)

Runs Test
data: res
statistic = 0,32036, runs = 22, n1 = 20, n2 = 20, n = 40, p-value = 0,7487
alternative hypothesis: nonrandomness
```

Din rezultatele prezentate mai sus, se observă că semnificaţia testului este mai mare decât pragul de *0,05* (*p-value=0,74*). În concluzie, nu se respinge ipoteza nulă, adică se poate considera că erorile modelului sunt independente.

4. Corectarea autocorelării erorilor

Corectarea modelului pentru care se încalcă ipoteza de independenţă a erorilor se realizează în funcţie de următoarele două situaţii: este cunoscut coeficientul de corelaţie de ordinul întâi a erorilor şi nu se cunoaşte acest coeficient.

Se consideră procedeul de corecţie pentru modelul de regresie liniară simplă:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

a. Cazul ρ cunoscut

Pentru corectarea modelului, se utilizează modelul de *quasi-diferenţă*, adică modelul de regresie:

$y_i^* = \beta_0^* + \beta_1^* x_i^* + u_i$, unde

$\beta_0^* = \beta_0(1-\rho)$;

$\beta_1^* = \beta_1$;

$y_i^* = y_i - \rho y_{i-1}$;

$x_i^* = x_i - \rho x_{i-1}$;

$u_i = \varepsilon_i - \rho \varepsilon_{i-1}$.

Aplicând metoda celor mai mici pătrate pentru modelul de *quasi-diferență*, se obțin doi estimatori nedeplasați, convergenți și eficienți, adică $\hat{\beta}_0^*, \hat{\beta}_1^*$. Pe baza acestora, se obțin estimatorii pentru modelul inițial:

$$\hat{\beta}_0 = \frac{\hat{\beta}_0^*}{1-\rho},$$

$\hat{\beta}_1 = \hat{\beta}_1^*$.

Dacă nu există autocorelare, estimatorii sunt identici; dacă există autocorelare a erorilor, parametrul β_0 este estimat eficient de estimatorul $\hat{\beta}_0^*$. Cunoscând coeficientul de corelație a erorilor, se pot obține estimațiile parametrilor, pe baza datelor disponibile, utilizând relațiile de mai sus.

b. Cazul ρ necunoscut

Dacă nu este cunoscut coeficientul de corelație dintre erori, soluția este estimarea acestuia pe baza datelor de sondaj. O metodă de corectare a autocorelării erorilor care se bazează pe estimațiile coeficienților de corelație a erorilor este procedeul iterativ Cochrane - Orcutt.

Procedeul presupune parcurgerea următorului demers:

- se construiește modelul de regresie:
  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$;
- se estimează erorile modelului și se obțin valorile e_i;
- pe baza erorilor estimate, se construiește modelul:
  $\varepsilon_i = \rho \varepsilon_{i-1} + u_i$
  și se estimează punctual parametrul ρ, adică se obține o estimație a acestuia, respectiv *r*. Estimația din prima iterație se notează cu $r^{(1)}$;
- se construiește modelul de *quasi-diferență*:
  $y_i^* = \beta_0^* + \beta_1^* x_i^* + u_i$,
  cu ajutorul estimației parametrului, determinate la pasul 3;
- se reia pasul 1 cu rezultatele de la pasul 4, dacă modelul obținut în prima iterație este în continuare influențat de autocorelare. Procedeul continuă cu o nouă iterație, care are ca punct de plecare modelul obținut în prima iterație.

BCU CLUJ CENTRALA UNIVERSITARĂ LIBRARY

În cea de-a doua iteraţie, se estimează un alt coeficient de autocorelare a erorilor pe care îl notăm $r^{(2)}$.

Procedeul se opreşte atunci când între două valori estimate ale coeficientului de autocorelaţie din două iteraţii succesive se verifică relaţia:

$r^{(p)} - r^{(p-1)} = 0{,}0025$.

7.2. Ipoteze asupra variabilelor independente

Această secţiune are în vedere ipotezele ce privesc variabilele independente ale unui model de regresie. Mai întâi, sunt prezentate aceste ipoteze, iar apoi sunt prezentate detalii cu privire la situaţia de multicoliniaritate.

7.2.1. Prezentare ipoteze

Pentru variabilele independente, sunt valabile mai multe ipoteze care funcţionează ca restricţii de modelare.

O primă restricţie este legată de gradul de omogenitate a variabilelor independente. Deoarece în relaţiile varianţelor estimatorilor apare varianţa variabilelor independente, este important ca această varianţă să fie posibil de calculat, să fie finită şi diferită de zero.

O altă ipoteză este condiţia ca variabilele independente să nu fie corelate cu variabilele reziduale. Această restricţie este respectată dacă este îndeplinită condiţia ca variabilele independente să fie variabile deterministe sau nealeatoare.

Cea mai importantă ipoteză asupra variabilelor independente este cea de necoliniaritate, care va fi tratată separat în continuare.

7.2.2. Ipoteza de necoliniaritate a variabilelor independente

1. Definirea ipotezei

Ipoteza este valabilă pentru modelele de regresie liniară multiplă, care au două sau mai multe variabile independente. Condiţia impusă de această ipoteză este ca între variabilele independente să nu existe o legătură de tip liniar.

În cazul existenţei coliniarităţii, se impune identificarea gradului de coliniaritate. Pentru un model de regresie care are p variabile independente, se definesc două tipuri de coliniaritate: perfectă şi imperfectă.

Între variabilele independente există o coliniaritate *perfectă* dacă există p constante $\lambda_1, \lambda_2, ..., \lambda_p$, nu toate nule, astfel încât să aibă loc relaţia:

$$\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_p X_p = 0.$$

Analog, între variabile există o coliniaritate *imperfectă* dacă pentru p constante $\lambda_1, \lambda_2, ..., \lambda_p$, nu toate nule, are loc relaţia:

$$\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_p X_p + u = 0,$$

unde u este o variabilă pur aleatoare, adică respectă ipotezele pentru componenta aleatoare a unui model de regresie.

Coliniaritatea poate apărea din mai multe surse: tipul de model de regresie utilizat, natura fenomenului și variabilele alese pentru a realiza modelarea etc. Este important de precizat că fenomenul apare la nivelul eșantionului de date disponibile, în contextul estimării parametrilor modelului, și nu la nivelul populației totale.

2. Efectele coliniarității

Dacă pentru un model de regresie multiplă, variabilele independente sunt coliniare, varianța estimatorilor parametrilor modelului de regresie crește, adică estimatorii pierd proprietatea de eficiență. Dacă se înregistrează o coliniaritate perfectă, varianța estimatorilor este infinită, ceea ce înseamnă că parametrii pentru aceste variabile independente nu pot fi estimați. Dacă se înregistrează o coliniaritate imperfectă, variațele estimatorilor pentru parametrii modelului de regresie sunt mari.

3. Testarea coliniarității

Un prim indiciu pentru existența coliniarității poate fi următorul: dacă între variabilele independente există o legătură de tip liniar, cel mai probabil, coeficientul de determinație pentru acest model va avea o valoare ridicată, însă testul Student pentru fiecare parametru al variabilelor coliniare nu va fi semnificativ statistic.

În consecință, se poate testa coliniaritatea prin testarea coeficienților de regresie, iar indiciul este existența unui coeficient de determinație mare. În condițiile în care parametrii modelului de regresie sunt nesemnificativi, se poate decide că modelul admite fenomenul de coliniaritate.

O altă metodă de testare a coliniarității este testarea parametrilor modelelor de regresie auxiliară construite ca modele de regresie liniară doar pe baza variabilelor independente. Dacă parametrii acestor modele sunt semnificativi, atunci variabilele independente sunt coliniare.

Pe baza modelelor de regresie auxiliare, se pot construi doi indicatori cu ajutorul cărora se poate detecta existența coliniarității. În soft-urile de statistică, acești indicatori sunt denumiți *Tolerance* și *VIF (Variance Inflation Factor)*.

Considerăm un model de regresie multiplă cu două variabile independente:

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$.

Pentru acest model, variațele estimatorilor parametrilor sunt:

$$V(\hat{\beta}_1)=\frac{\sigma^2}{\sum_i (x_{1i}-\bar{x}_1)^2(1-R_{12}^2)},$$

$$V(\hat{\beta}_2)=\frac{\sigma^2}{\sum_i (x_{2i}-\bar{x}_2)^2(1-R_{12}^2)}, \text{ unde}$$

$$R_{12}^2=\frac{\left[\sum_i (x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)\right]^2}{\left[\sum_i (x_{1i}-\bar{x}_1)^2\right]\left[\sum_i (x_{2i}-\bar{x}_2)^2\right]}.$$

Ultima relaţie prezintă raportul de determinaţie dintre variabilele independente din modelul de regresie auxiliar.

Pentru cazul unui model de regresie cu p variabile independente, varianţa estimatorului parametrului $\beta_j, j=\overline{1,p}$, asociat variabilei independente X_j, este:

$$V(\hat{\beta}_j)=\frac{\sigma^2}{\sum_i (x_{ji}-\bar{x}_j)^2(1-R_j^2)},$$

unde R_j^2 este raportul de determinaţie din modelul de regresie auxiliar, construit pe baza variabilelor independente. În acest model, variabila X_j este variabila dependentă, iar celelalte variabile factoriale sunt variabile independente.

Pentru p variabile independente, modelul auxiliar se poate scrie astfel:

$$X_j=\alpha_0+\alpha_1X_1+...+\alpha_{j-1}X_{j-1}+\alpha_{j+1}X_{j+1}+...+\alpha_pX_p+u\ .$$

Indicatorul *VIF* se defineşte prin relaţia:

$$VIF_j=\frac{1}{(1-R_j^2)}.$$

Acesta indică modul în care varianţa estimatorului unui coeficient de regresie este influenţată de prezenţa coliniarităţii la nivelul variabilelor independente.

Interpretare

- valoarea $VIF_j=1$ indică lipsa coliniarităţii şi se realizează atunci când $R_j^2=0$. Dacă $R_j^2=1$, între variabilele independente există o coliniaritate perfectă, iar valoarea VIF_j este infinită.

BCU "SPIRU HARET" UNIVERSITY LIBRARY

- dacă variabilele sunt coliniare, indicatorul VIF_j are o valoare ridicată. În practică, pentru o valoare $VIF_j > 10$ (alți autori propun un $VIF_j > 5$), se consideră că este prezent fenomenul de coliniaritate.

Indicatorul *Tolerance* se determină ca inversul indicatorului *VIF*:

$$TOL_j = \frac{1}{VIF_j} = (1 - R_j^2).$$

Interpretare

Pentru $TOL_j = 1$, variabilele independente nu sunt coliniare, iar dacă $TOL_j = 0$, există coliniaritate perfectă. Existența coliniarității este sugerată de valorile mici ale indicatorului TOL_j.

4. Corectarea coliniarității

Metodele de corecție a coliniarității trebuie să țină cont de tipul de coliniaritate dintre variabile, de numărul de variabile din model și de informațiile suplimentare despre fenomenul studiat.

În literatura de specialitate, se întâlnesc mai multe metode de corectare a coliniarității.

Cea mai facilă metodă este eliminarea variabilei care introduce coliniaritatea la nivelul modelului de regresie. În această situație însă, există riscul eliminării din model a unei variabile importante pentru explicarea fenomenului studiat.

O altă metodă este construirea unui model de regresie cu variabile transformate prin diverse funcții sau operatori (de exemplu, prin operatorul decalaj, diferență), iar în acest mod, se poate elimina dependența liniară dintre variabilele factoriale.

Exemplu

La nivelul județelor României, pentru variabilele *Rata brută de natalitate* (*Y*, născuți vii la 1000 de locuitori), *Rata brută de nupțialitate* (X_1, căsătorii la 1000 de locuitori), *Rata brută de divorțialitate* (X_2, divorțuri la 1000 de locuitori) și *Rata mortalității infantile* (X_3, decese sub 1 an la 1000 de născuți vii), s-a construit un model de regresie multiplă pe baza datelor din anul 2015.

Pentru aceste variabile, se estimează un model de regresie liniară multiplă. Testarea multicoliniarității în R presupune calculul indicatorului *VIF* cu ajutorul comenzii:

```
> vif(lm(Nat~Cas+Div+Inf.))
     Cas      Div      Inf
1,555267 1,038323 1,525219
```

Interpretarea indicatorilor de coliniaritate

Valoarea indicatorului *VIF* este mică pentru toate cele 3 variabile independente (mai mică de limita egală cu 5), ceea ce indică lipsa coliniarității dintre variabilele independente utilizate în model.

Observație

Dispunând de valoarea *VIF*, se poate estima valoarea coeficientului de determinație din modelul auxiliar corespunzător fiecărei variabile independente și se poate analiza mărimea acestei valori.

Pentru exemplul de mai sus, valorii $VIF_1 = 1,55$ îi corespunde o valoare a coeficientului de determinație: $R_1^2 = 0,35$.

7.3. Teste și exerciții

Teste grilă

1. Un model de regresie este homoscedastic dacă:
a) erorile modelului au media egală cu zero;
b) erorile de modelare sunt independente;
c) varianțele erorilor de modelare sunt egale;
d) erorile au dispersia cuprinsă în intervalul (0,1).

2. Testul Durbin-Watson se utilizează pentru testarea:
a) coliniarității variabilelor factoriale;
b) homoscedasticității erorilor;
c) independenței erorilor.

3. Într-un model de regresie liniară multiplă, dacă variabilele independente sunt perfect coliniare:
a) dispersia estimatorilor parametrilor este zero;
b) dispersia estimatorilor parametrilor este infinită;
c) erorile de modelare sunt minime;
d) parametrii modelului nu pot fi estimați.

4. Testul Goldfeld-Quandt:
a) este utilizat pentru verificarea ipotezei de homoscedasticitate;
b) are la bază o statistică Fisher;
c) este util pentru verificarea independenței erorilor.

5. În testarea autocorelării erorilor, dacă valoarea calculată a statisticii Durbin-Watson este $d = 0$, se poate considera că:

a) erorile sunt dependente;
b) există autocorelare negativă maximă între erori;
c) nu există autocorelare între erori;
d) există autocorelare pozitivă maximă între erori.

6. Pentru testarea ipotezei de normalitate a erorilor se utilizează testele:
a) QQPlot;
b) Student;
c) Jarque-Bera;
d) Fisher.

7. În vederea testării ipotezei de necorelare a erorilor unui model de regresie liniară, s-au obținut următoarele rezultate:

```
> dwt(res)
[1] 1,43
```

Cunoscând valorile critice din tabela Durbin-Watson $d_L = 1,503$ și $d_U = 1,585$, pentru un risc de 0,05, se poate considera că:
a) erorile de modelare sunt autocorelate pozitiv;
b) erorile de modelare sunt autocorelate negativ;
c) nu este posibilă luarea unei decizii cu privire la existența autocorelării erorilor.

8. În vederea testării ipotezei de normalitate a erorilor unui model de regresie liniară simplă, prin prelucrarea datelor pentru un eșantion de volum $n = 11$ unități, s-au obținut următoarele rezultate:

```
> skewness(res)
[1] -0,25
> kurtosis(res)
[1] 1,06
```

Cunoscând valoarea teoretică a statisticii test, $\chi^2_{0,05;2} = 5,99$, se poate considera că:
a) erorile de modelare urmează o lege de repartiție normală;
b) erorile de modelare nu urmează o lege de repartiție normală;
c) erorile de modelare sunt independente.

9. Pentru un model de regresie liniară multiplă, s-au obținut următoarele rezultate:

BCU "ION CREANGĂ" UNIVERSITY LIBRARY

> vif(lm(Y~X_1+X_2+X_3))

X_1	X_2	X_3
2,4	5,1	11,2

Se poate concluziona că:

a) variabilele independente nu sunt coliniare;

b) două dintre variabile respecta ipoteza de necoliniaritate;

c) variabilele independente nu sunt coliniare.

10. Dacă pentru un model de regresie liniară multiplă indicatorul *Tolerance* ia valoarea $TOL = 1$, atunci variabilele independente sunt:

a) coliniare;

b) necoliniare;

c) dependente.

Exerciții

1. Comentați diferența dintre cele două relații de mai jos:

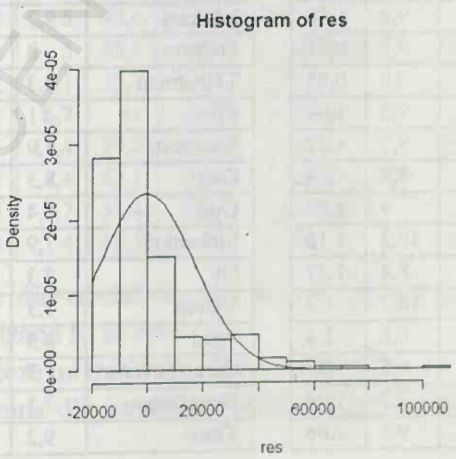
$\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_p X_p = 0$;

$\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_p X_p + \varepsilon = 0$.

2. Se poate vorbi despre multicoliniaritate într-un model polinomial de forma: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$?

Cum gestionăm situația?

3. Histograma erorilor estimate pentru un model de regresie multiplă, la nivelul unui eșantion de 500 de unități se prezintă mai jos.



BCU IAȘI CENTRAL UNIVERSITY LIBRARY

Comentați graficul de mai sus în privința formei distribuției. Se pot face aprecieri asupra ipotezei de normalitate a erorilor?

4. Pentru un model de regresie cu 2 variabile independente, s-a aplicat metodologia cerută de testul Goldfeld-Quandt și s-au construit 2 subeșantioane de date de volum egal cu 100. Pentru acestea s-au obținut următoarele rezultate:

$RSS_1 = 3245$ și $RSS_2 = 12450$.

Să se verifice ipoteza de homoscedasticitate pentru cazul considerat.

5. Să reluăm discuția din capitolul 5 privind repartiția angajaților după nivelul de educație și salariu. Propunem două modele de analiză a legăturii dintre cele două variabile:

$Y = \beta_0 + \beta_1 X + \varepsilon$;

$\ln Y = \beta_0 + \beta_1 X + \varepsilon$.

Pentru care dintre cele două modele sunt șanse mai mari să existe heteroscedasticitate? Motivați răspunsul.

6. Se dau datele din tabelul de mai jos privind ratele de natalitate și divorțialitate, la nivelul județelor României, în anul 2014 (sursa www.insse.ro).

Județul	Nat.	Div.
Bihor	9,3	1,2
Bistrita-Nasaud	10	0,6
Cluj	9,1	0,97
Maramures	8,7	1,32
Satu Mare	9,1	1,26
Salaj	10,1	1,1
Alba	8,2	1,34
Brasov	9,8	1,78
Covasna	9,8	0,71
Harghita	10	0,85
Mures	9,5	1,04
Sibiu	9,7	1,22
Bacau	8,8	1,34
Botosani	9	1,62
Iasi	10,3	1,16
Neamt	8,4	1,37
Suceava	10,7	1,3
Vaslui	8,8	1,4
Braila	6,5	1,95
Buzau	7,7	1,09
Constanta	9,5	1,66

Județul	Nat.	Div.
Galati	7,8	1,18
Tulcea	8,1	0,83
Vrancea	8,8	0,91
Arges	8	1,39
Calarasi	8,5	1,29
Dâmbovita	8,4	1,03
Giurgiu	8,2	0,53
Ialomita	9,2	0,91
Prahova	8	1,33
Teleorman	7,3	1,13
Ilfov	11	1,28
Bucuresti	8,9	1,31
Dolj	8,3	0,63
Gorj	7,4	1,55
Mehedinti	7,9	0,85
Olt	7,3	0,93
Vâlcea	7,3	0,98
Arad	8,5	1,12
Caras-Severin	7	1,66
Hunedoara	7,1	1,55
Timis	9,2	1,1

BCU IAȘI - BIBLIOTECA UNIVERSITARĂ

Se cere:

- să se estimeze un model liniar simplu pentru cele 2 variabile cu ajutorul programului R;
- verificați ipoteza de homoscedasticitate cu ajutorul a două teste din cele cunoscute;
- în cazul în care nu se respectă condiția de homoscedasticitate, realizați o corecție a modelului folosind o metodă cunoscută;
- verificați ipoteza de normalitate a erorilor folosind un test grafic și un test numeric;
- ce se întâmplă dacă ipoteza de normalitate nu este respectată?

7. La nivelul județelor României, în anul 2014, considerăm variabilele prezentate în tabelul de mai jos: PIB pe locuitor (lei), rata șomajului (%), gradul de urbanizare (%), ponderea populației vârstnice (%).

Județul	pib	som	urb	aged
Bihor	26683	3,7	49,2	15,6
Bistrita-Nas,	23870	4,5	37,6	15,5
Cluj	42579	2,8	66,1	16,4
Maramures	23080	3,5	57,3	15,2
Satu Mare	23774	4,4	45,1	14,1
Salaj	26213	5,9	39,5	18,1
Alba	32669	7,3	58,2	17,9
Brasov	39954	4,3	71,6	15,2
Covasna	22710	6,5	47,5	15,6
Harghita	22917	6	42,6	16,1
Mures	26409	5,8	49,9	17
Sibiu	34769	4,6	65,9	14,8
Bacau	21182	6,8	43,1	17
Botosani	16533	5	40,6	17,8
Iasi	26281	5	45,9	14,2
Neamt	19182	6,1	35,7	19,3
Suceava	19413	6,7	41	16,3
Vaslui	15441	11,4	39,4	17,3
Braila	23733	7,4	62,1	19,5
Buzau	21483	9,7	38,4	19,8
Constanta	49564	4	68,3	14,6

Județul	pib	som	urb	aged
Galati	24048	9,5	54,6	16,4
Tulcea	24887	5,5	46,6	16,6
Vrancea	20146	6	36,2	18,7
Arges	29698	5,9	45,9	17,1
Calarasi	19485	8,2	36,1	18,6
Dâmbovita	22690	7,9	28,6	16,5
Giurgiu	21683	7	29,1	19,8
Ialomita	22990	8,1	44	17,9
Prahova	43470	5,2	49	18,1
Teleorman	18419	11,2	32,3	24,4
Ilfov	40447	1,5	43,7	12,7
Bucuresti	87021	2	100	15
Dolj	25521	9,4	51,9	18,3
Gorj	29972	7,5	45,2	16
Mehedinti	18404	10	46,3	18,6
Olt	20094	8,2	39	19,6
Vâlcea	23530	5,6	44,6	20,3
Arad	32365	2,8	55,3	16,9
Caras-Severin	23722	4,5	53,9	18
Hunedoara	26889	6,6	74,7	18,1
Timis	42897	1,6	61,1	14,1

Pentru datele din tabelul de mai sus, se consideră ca variabilă dependentă rata șomajului. Utilizând R, se cere:

a) să se estimeze un model de regresie multiplă;

b) să se elimine din model variabilele nesemnificative folosind procedura *Backward*;

c) pentru modelul cu variabile semnificative, verificați ipoteza de homoscedasticitate cu ajutorul testului Breusch-Pagan;

d) similar, verificați ipoteza de normalitate a erorilor, atât pe cale grafică, precum și cu ajutorul testului Jarque-Bera;

e) în cazul ipotezei de multicoliniaritate, să se calculeze coeficienții VIF_j pentru variabilele independente și să se compare, pe rând, cu valoarea 5, respectiv 10. La ce concluzie se ajunge?

BCU IAȘI/CENTRAL UNIVERSITY LIBRARY

Capitolul 8

MODELAREA SERIILOR DE TIMP

În capitolele anterioare au fost prezentate elemente metodologice privind modelarea datelor economice și sociale pentru populații sau eșantioane de date de moment (*cross-section data*). Pentru modelarea seriilor de timp, se urmează aproximativ același demers, însă există o serie de elemente specifice. În general, pentru modelarea seriilor de timp există lucrări separate care fac obiectul unor discipline precum *Econometria seriilor de timp*. În lucrările clasice de econometrie, pentru seriile de timp, sunt prezentate succint câteva elemente de bază cu privire la: definirea unei serii de timp ca un proces stocastic, staționaritatea unui proces, modelarea unei serii ținând cont de componentele acesteia; prognoza unei serii de timp etc. În limitele acestei lucrări, în acest capitol, se prezintă câteva aspecte ale acestei problematici.

8.1. Seriile de timp ca procese stocastice

Modelarea econometrică a unei serii de timp presupune o abordare specifică, în comparație cu tratarea acestora în analiza statistică clasică.

8.1.1. De la abordarea statistică la abordarea econometrică a unei serii de timp

1. Abordarea statistică

Din punct de vedere statistic, o serie de timp reprezintă un set de observații sau măsurători înregistrate în timp, pentru un anumit fenomen. În egală măsură, putem vorbi despre un set de valori pe care le ia o variabilă (cantitativă sau calitativă) la diferite momente sau intervale de timp. Pentru o serie de timp, pe lângă modul de exprimare (de moment și de interval), este important să se

BCU "ION CREANGĂ" UNIVERSITY LIBRARY

cunoască periodicitatea sau frecvența cu care apar datele: anuală, semestrială, trimestrială, lunară, săptămânală, zilnică etc. Este important ca pentru o anumită variabilă, datele înregistrate să aibă aceeași frecvență.

Analizei statistice a unei serii de timp îi este specific studiul componentelor: trend, sezonieră, ciclică, aleatoare. De exemplu, în figura 8.1, se prezintă evoluția unui fenomen în timp, iar forma grafică ne permite identificarea componentelor seriei:

- pe termen lung, seria înregistrează o tendință de creștere – așadar, putem vorbi despre componenta *trend*;
- pe termen mediu, pot exista anumite variații cu caracter regulat care să se constituie ca o componentă distinctă – componenta *ciclică*;
- pe termen scurt, se pot observa variații regulate cu caracter repetitiv – componenta *sezonieră*;
- de la un moment de timp la altul, se observă variații neregulate – componenta *aleatoare*.

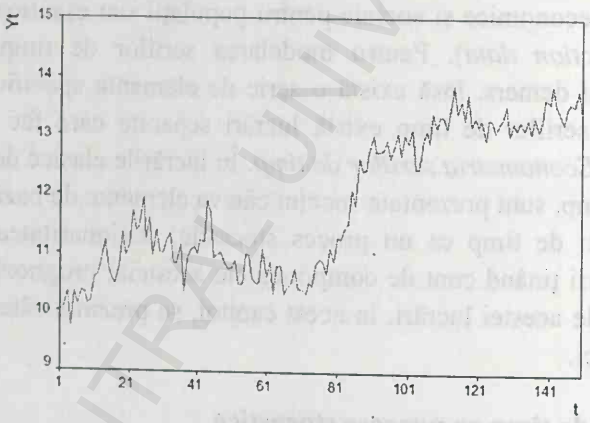


Figura 8.1. *Exemplu de serie de timp*
Sursa: Realizată de autor (date convenționale)

Observație

Identificarea vizuală a anumitor componente nu este suficientă. Analiza seriilor de timp presupune utilizarea unor metode care să capteze aceste variații și să le instrumenteze. De asemenea, analiza statistică poate dovedi dacă anumite componente există sau nu, deoarece pentru o serie de timp nu este necesar să existe toate cele patru componente.

2. Abordarea econometrică

În analiza econometrică, o serie de timp este tratată ca un proces stocastic, iar un set concret de date (o serie de timp statistică) reprezintă o realizare a

procesului. Distincţia dintre proces stocastic şi o realizare a acestuia se poate face în mod analog cu distincţia dintre populaţie şi eşantion din statistica inferenţială.

Procesul stocastic sau procesul aleator reprezintă instrumentul (modelul) econometric de analiză a unui fenomen care se realizează în timp şi a cărui evoluţie este supusă acţiunii factorului aleator.

Factorul aleator care acţionează în evoluţia fenomenului poate fi privit ca o influenţă perturbatoare asupra unei legi de evoluţie a fenomenului sau poate constitui însăşi natura evoluţiei fenomenului, aşa cum este cazul unui proces bine cunoscut în literatură – procesul „zgomot alb".

Analiza seriei de timp, ca proces stocastic, presupune studiul caracteristicilor numerice ale acestui instrument: speranţa matematică sau media, varianţa, covarianţa. Aceste caracteristici pot varia în timp sau pot fi constante.

8.1.2. Procese stocastice. Definiţie şi caracteristici

Această secţiune este dedicată analizei instrumentului statistic de bază în modelarea seriilor de timp: procesul stocastic. Pe lângă definiţie şi o serie de caracteristici, este prezentat un exemplu de proces stocastic frecvent utilizat în analiza econometrică.

1. Definiţie

Procesul stocastic este definit de o familie de variabile aleatoare care depinde de variabila timp. Formal, procesul stocastic este dat prin expresia:

$\{Y_t, t \in T\}$.

Variabila Y_t este o variabilă aleatoare şi are o densitate de probabilitate notată cu $f(y_t)$ şi este numită *funcţie densitate necondiţionată a lui* Y_t.

Prin $(y_1, y_2, \ldots, y_T)$, se notează o posibilă realizare a procesului stocastic. Acest set de date reprezintă T valori particulare dintr-un ansamblu (infinit) de valori ale unei realizări posibile a procesului:

$\{\ldots, y_{-1}, y_0, y_1, y_2, \ldots, y_{T-1}, y_T, y_{T+1}, \ldots\}$.

În general, pentru n posibile realizări ale procesului stocastic, se poate considera un eşantion de forma:

$\{y_t^{(1)}, y_t^{(2)}, \ldots, y_t^{(n)}\}$,

adică este un eşantion de n realizări ale variabilei aleatoare Y_t.

Observaţie

În R, o serie de timp se defineşte cu ajutorul funcţiei *ts* (în pachetul *tseries*):

```
> ts(variabile, start =, end =, frequency =)
```

BCU IASI / CENTRAL UNIVERSITY LIBRARY

Opțiunile *start* și *end* se referă la momentele de timp pentru prima și ultima observație din serie, iar frecvența vizează numărul de observații pe unitatea de timp (1 pentru anual, 4 pentru trimestrial, 12 pentru lunar). De exemplu, dacă seria începe în octombrie 1990, se utilizează opțiunea *start = c(1990,10)*.

2. Operatori utilizați în modelarea seriilor de timp

Analiza econometrică a seriilor de timp presupune frecvent transformarea datelor utilizate cu ajutorul unor operatori matematici. Dintre cei mai utilizați operatori, amintim: decalaj, diferență, log-diferență.

a. Operatorul decalaj (lag)

Acest operator realizează corespondența dintre valoarea seriei la momentul t și un moment anterior.

Operatorul de ordinul întâi este definit de relația:

$LY_t = Y_{t-1}$.

Operatorul de ordin k este:

$L^{(k)}Y_t = Y_{t-k}$.

Proprietățile acestui operator sunt:

$L(LY_t) = LY_{t-1} = Y_{t-2}$;

$L(\beta Y_t) = \beta LY_t = \beta Y_{t-1}$;

$L(Y_t + \varepsilon_t) = LY_t + L\varepsilon_t = Y_{t-1} + \varepsilon_{t-1}$.

b. Operatorul diferență

Operatorul diferență realizează corespondența dintre valoarea seriei la momentul t și diferența dintre valoarea seriei la momentul t și valoarea seriei la un moment de timp anterior.

Operatorul diferență de ordinul întâi este:

$DY_t = Y_t - Y_{t-1}$.

Pentru un ordin k, operatorul are relația:

$D^{(k)}Y_t = DD^{(k-1)}Y_t$.

De exemplu, pentru $k = 2$, are loc relația:

$D^{(2)}Y_t = D(Y_t - Y_{t-1}) = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$.

Dacă se ține cont de operatorul decalaj, operatorul diferență este:

$DY_t = Y_t - Y_{t-1} = Y_t - LY_t = (1-L)Y_t$ sau $D = 1 - L$;

$D^{(2)}Y_t = (1-L)^2 Y_t = (1 - 2L + L^{(2)})Y_t$;

$D^{(k)}Y_t = (1-L)^k Y_t$.

BCU "ION IONESCU DE LA BRAD"

Operatorul diferenţă este utilizat în procedeele de staţionarizare a unei serii de timp.

c. Operatorul log-diferenţă

Acest operator este frecvent utilizat pentru a construi rate de creştere, pentru un anumit fenomen. Relaţia de definiţie a operatorului de ordinul 1 este:

$$D\log Y_t = \log Y_t - \log Y_{t-1}.$$

Acest operator este echivalent cu logaritmul raportului dintre cele două variabile:

$$D\log Y_t = \log\frac{Y_t}{Y_{t-1}}.$$

Observaţie

În analiza seriilor de timp, prin aplicarea operatorilor de mai sus, se construiesc două variabile importante:

- rata de creştere a unei variabile: $R_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}}100\%$;
- rata de creştere continuă: $r_t = D\log Y_t = \log\frac{Y_t}{Y_{t-1}}100\%$.

3. Caracteristici numerice ale unui proces stocastic

Pentru un proces stocastic, se determină în principal următoarele mărimi: media sau speranţa matematică, varianţa, covarianţa sau autocovarianţa. Pe baza acestora, se mai determină funcţia de autocorelaţie şi funcţia de autocorelaţie parţială, care iau valori concrete pentru un anumit decalaj. Aceste caracteristici se determină pe legea de repartiţie a procesului, care este determinată de o familie de repartiţii de probabilitate.

Speranţa matematică

Media procesului stocastic este notată cu $\mu_t = E(Y_t)$, $t \in T$, şi se determină distinct, în funcţie de tipul procesului:

$E(Y_t) = \int_R y_t f(y_t)\, dy_t$, pentru procesul continuu;

$E(Y_t) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} Y_t^{(i)}$, pentru procesul discret.

Varianţa

Varianţa procesului stocastic se notează cu γ_{0t} şi este definită prin relaţia:

$$\gamma_{0t} = E(Y_t - \mu_t)^2.$$

B.C.U. „M. EMINESCU" IAŞI

Covarianța sau funcția de autocovarianță

Covarianța la un decalaj k reprezintă covarianța dintre variabilele Y_t și Y_{t-k}:

$\gamma_{kt} = E(Y_t - \mu_t)(Y_{t-k} - \mu_{t-k})$.

Pentru un decalaj $k = 0$, se obține varianța procesului stocastic:

$\gamma_{0t} = E(Y_t - \mu_t)^2$.

Funcția de autocorelație

Funcția de autocorelație la un decalaj k se definește prin relația:

$$\rho_k = \frac{\gamma_{kt}}{\gamma_{0t}}.$$

Această mărime măsoară intensitatea legăturii dintre variabilele Y_t și Y_{t-k}. Ca orice coeficient de corelație, respectă condiția:

$-1 \le \rho_k \le 1$.

Funcția de autocorelație parțială

Această funcție se determină ca un coeficient de corelație parțială între variabila de la momentul t și cea de la momentul $t-k$. Coeficientul de corelație parțială măsoară intensitatea legăturii dintre cele două variabile, Y_t și Y_{t-k}, controlând influența variabilelor pentru un decalaj diferit.

4. Procesul stocastic pur sau „zgomotul alb"

În modelarea seriilor de timp, se utilizează un proces stocastic cu proprietăți specifice. Este vorba despre procesul compus dintr-o familie de variabile aleatoare, notată cu ε_t, care respectă toate ipotezele modelului clasic de regresie.

Procesul aleatoriu este dat prin relația:

$Y_t = \varepsilon_t$, unde $\varepsilon_t \sim N(0, \sigma^2)$ și are funcția densitate de probabilitate:

$$f(y_t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y_t^2}{2\sigma^2}}.$$

Acest proces poartă numele de *proces gaussian* sau „zgomot alb" și admite următoarele caracteristici numerice:

$E(Y_t) = 0$;

$V(Y_t) = \sigma^2$.

Grafic, un proces de acest tip este reprezentat în figura 8.2.

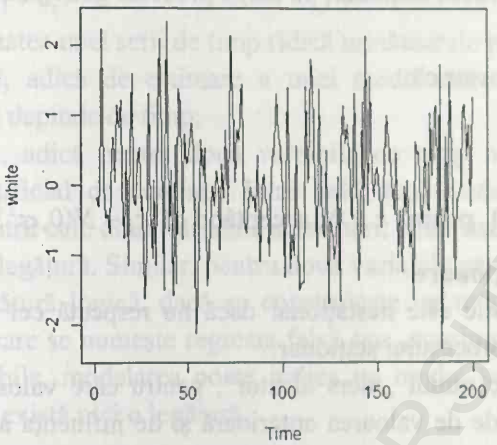


Figura 8.2. *Exemplu de proces de tip „zgomot alb"*
Sursa: Realizată de autor (date convenționale)

Procesul a fost generat în R cu ajutorul comenzilor:

```
> white=arima.sim(list(order=c(0,0,0)),200)
> plot(white)
```

8.1.3. Staționaritatea proceselor stocastice. Procese integrate

În modelarea seriilor de timp, cu ajutorul proceselor stocastice, una dintre trăsăturile importante pe care trebuie să le îndeplinească seria de timp este staționaritatea. În acest subcapitol, se definește proprietatea de staționaritate și se prezintă implicațiile pe care le are pentru modelare, precum și modalitățile de verificare a acestei proprietăți pentru o serie de timp.

1. Definiție

Un proces stocastic se numește staționar, în sens larg, dacă media și varianța sa nu depind de timp, adică sunt constante în timp, iar covarianța la un decalaj k nu depinde decât de acel decalaj (nu depinde de timp).

Pentru media procesului stocastic, condiția de staționaritate presupune ca speranța matematică a seriei să fie constantă în timp sau valoarea sperată a unei valori a seriei pentru un moment de timp viitor să fie aceeași ca pentru momentul prezent sau un moment din trecut.

Procesul stocastic este staționar dacă au loc condițiile:

$E(Y_t) = \mu$;

$\gamma_{0t} = E(Y_t - \mu_t)^2 = \sigma^2$;

$E(Y_t - \mu_t)(Y_{t-k} - \mu_{t-k}) = E(Y_t - \mu)(Y_{t-k} - \mu) = \gamma_k$.

BCU BUCUREȘTI UNIVERSITY LIBRARY

Un exemplu de proces staționar, pe lângă procesul gaussian, este:

$Y_t = \mu + \varepsilon_t$.

Se demonstrează ușor că:

$E(Y_t) = \mu$,

$\gamma_{0t} = \sigma^2$,

$\gamma_k = E(\varepsilon_t \varepsilon_{t-k}) = 0$, pentru $k \neq 0$, acceptând că $\varepsilon_t \sim N(0, \sigma^2)$.

2. Procese nestaționare

Un proces stocastic este nestaționar dacă nu respectă cel puțin una dintre condițiile de definire a procesului staționar.

Procesul stocastic numit „mers aleator", pentru care valoarea variabilei la momentul curent depinde de valoarea anterioară și de influența aleatoare, este un proces nestaționar.

Acest proces este definit prin relația:

$Y_t = \mu + Y_{t-1} + \varepsilon_t$.

Dacă au loc relațiile:

$\varepsilon_t \sim N(0, \sigma^2)$,

$E(\varepsilon_t \varepsilon_{t-k}) = 0$,

atunci modelul respectă proprietățile:

$E(Y_t) = E(\sum \varepsilon_t) = t \cdot \mu = \mu_t$, iar

$V(Y_t) = V(\sum \varepsilon_t) = t \cdot \sigma^2$.

3. Probleme ale modelării proceselor stocastice

Pentru modelele econometrice, ipotezele de modelare se referă la procese staționare, adică la procese pentru care: media, estimată prin media empirică, este constantă, nu depinde de timp; varianța procesului, estimată prin varianța empirică, este constantă, nu depinde de timp; autocovariațele procesului, covariațele dintre un moment prezent și momentele de timp trecute, nu depind de timp.

Aceste condiții sunt rareori respectate în practică. În cele mai numeroase cazuri, variabile economice care depind de timp sunt nestaționare.

Nestaționaritatea poate fi:

- deterministă, adică variabilele sunt funcții de timp. Un exemplu de asemenea model este:

  $Y_t = \mu + \beta t + \varepsilon_t$;

- stocastică, adică variabilele sunt modelate de procese autoregresive. Un exemplu este modelul:

- $Y_t = \mu + \rho Y_{t-1} + \varepsilon_t$.

Nestaţionaritatea unei serii de timp ridică următoarele probleme:

- de *inferenţă*, adică de estimare a unei medii sau varianţe care nu este constantă, ci depinde de timp;
- de *corelaţie*, adică pentru două variabile de timp nestaţionare, apare un fenomen artificial de corelaţie între cele două variabile. Coeficienţii de corelaţie pentru cele două variabile sunt mari, chiar dacă între variabile nu ar exista nicio legătură. Similar, pentru două variabile nestaţionare între care nu există o legătură logică, dacă se construieşte un model de regresie, apare fenomenul care se numeşte regresia falsă sau „spurious regressions". Pentru aceste variabile, modelarea poate indica un model semnificativ, deşi între variabile nu există nici o legătură.

4. Testarea staţionarităţii unei serii de timp

În procesul de modelare econometrică a seriilor de timp, o etapă este verificarea staţionarităţii seriei de timp. Dacă seria este staţionară, modelarea se realizează în condiţiile clasice cunoscute. Dacă seria este nestaţionară, este important să se identifice tipul de nestaţionaritate şi apoi să se aplice metode de staţionarizare, pentru a se putea realiza modelarea.

În econometrie, se utilizează mai multe metode pentru testarea staţionarităţii unei serii de timp: metode grafice şi metode numerice. Metodele grafice au la bază coeficienţii de autocorelaţie pentru un număr de decalaje, iar metodele numerice se bazează pe teste.

a. Metoda grafică. Corelograma funcţiei de autocorelaţie

Această metodă presupune reprezentarea grafică a coeficienţilor de autocorelaţie până la un decalaj k. Valorile coeficienţilor se compară vizual cu cei ai unui proces aleator pur, cum este procesul gaussian.

Pentru un proces pur aleator (zgomot alb), funcţia de autocorelaţie are o distribuţie aproximativ normală cu medie zero şi varianţă $1/n$:

$\hat{\rho}_k \sim N(0,1/n)$.

Pentru o încredere de 95%, valorile funcţiei de autocorelaţie ale procesului gaussian respectă condiţia:

$$-1{,}96\frac{1}{\sqrt{n}} \le \hat{\rho}_k \le 1{,}96\frac{1}{\sqrt{n}}.$$

Pentru o serie de timp, pentru care se dispune de o realizare a procesului stocastic, metoda corelogramei presupune reprezentarea grafică a estimaţiilor funcţiei de autocorelaţie până la un decalaj k şi compararea vizuală a valorilor

acestora cu limitele intervalului $(-1{,}96\frac{1}{\sqrt{n}}; 1{,}96\frac{1}{\sqrt{n}})$. Dacă valorile funcției de autocorelație se păstrează în limitele intervalului de încredere, atunci seria de timp este staționară. De asemenea, dacă valorile funcției descresc exponențial, atunci seria este staționară.

Nestaționaritatea este indicată de cazul în care valorile coeficienților de autocorelație descresc lent și se află în afara intervalului de încredere.

Estimațiile coeficienților de corelație se determină pe baza estimatorului funcției de autocorelație pentru un decalaj k:

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}, \text{ unde:}$$

$$\hat{\gamma}_k = \frac{\sum(Y_t - \hat{\bar{Y}})(Y_{t-k} - \hat{\bar{Y}})}{n}, \quad \hat{\gamma}_0 = \frac{\sum(Y_t - \hat{\bar{Y}})^2}{n}.$$

Exemplu

Utilizăm baza de date oferită de Fondul Monetar Internațional, pentru perioada 1990-2012, privind indicele producției industriale (*output*, *Y*) și indicele prețurilor de consum (*inflation*, *X*). Baza de raportare este anul 2005, iar frecvența datelor este lunară.

Construim diagrama funcției de autocorelație în R (figura 8.3) pentru variabila *Y* cu ajutorul comenzii:

```
> acf(out, lag.max=20)
```

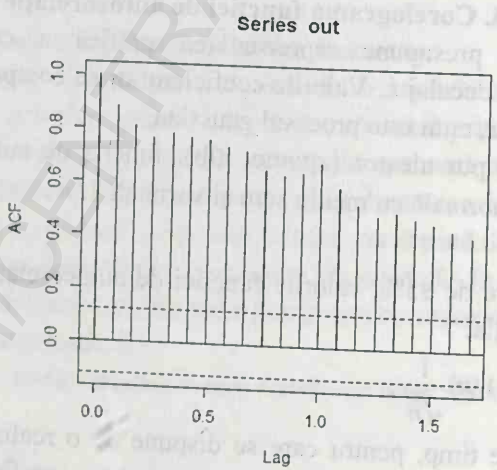


Figura 8.3. *Corelograma coeficienților de autocorelație*
Sursa: Realizată de autor pe baza datelor oferite de FMI, 2014

BCU IASI CENTRALA UNIVERSITARA

Analiza corelogramei conduce la concluzia că seria de timp nu este staţionară sau se abate semnificativ de la comportamentul specific unui proces de tip gaussian, deoarece valorile funcţiei de autocorelaţie descresc lent şi depăşesc limitele intervalului de încredere până la un decalaj egal cu 20.

b. Metode numerice

Pentru testarea staţionarităţii unei serii de timp, se utilizează mai multe teste: Ljung-Box, Dickey-Fuller, Phillips-Perron etc.

Testul Ljung-Box

Ipotezele pentru acest test sunt:

H_0: $\rho_k = 0$, toţi coeficienţii sunt simultan zero, seria este staţionară;

H_1: ρ_k nu sunt simultan zero, adică seria nu este staţionară.

Pentru testarea ipotezei nule, se utilizează statistica:

$$LB = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{n-k} \sim \chi^2(m).$$

Variabila *LB*, pentru un eşantion de volum n şi un decalaj m, urmează o lege de repartiţie Chi-pătrat de m grade de libertate.

Decizia se ia pe baza valorii calculate a testului, pentru o realizare a procesului stocastic, sau în funcţie de semnificaţia testului şi de pragul de semnificaţie stabilit. Ipoteza nulă (proces staţionar) este respinsă dacă valoarea calculată a testului este mai mare decât valoarea teoretică pentru un prag de semnificaţie stabilit şi m grade de libertate.

În R, se respinge ipoteza nulă dacă semnificaţia testului este mai mică decât pragul de semnificaţie ales.

Exemplu

Pentru seria de timp utilizată la metoda grafică, în R, se obţin următoarele rezultate cu privire la testul Ljung-Box, pentru un decalaj egal cu 1 şi apoi pentru un decalaj egal cu 10:

```
> Box.test(out, lag=1, type="Ljung-Box")

Box-Ljung test
data: out
X-squared = 206,96,   df = 1,   p-value < 2,2e-16

> Box.test(out, lag=10, type="Ljung-Box")

Box-Ljung test
data: out
X-squared = 1541,4,   df = 10,  p-value < 2,2e-16
```

Analiza rezultatelor de mai sus conduce la decizia de a nu accepta ipoteza nulă, deoarece semnificația testului este mai mică decât *0,05*.

Testul Dickey-Fuller (DF test)

Acest test[1] se mai numește și *testul rădăcinii unitate* și este important pentru că permite și identificarea tipului de nestaționaritate a seriei: stocastică sau deterministă.

Pentru testare, se consideră procesul $Y_t = \rho Y_{t-1} + \varepsilon_t$. Dacă coeficientul lui Y_{t-1} este egal cu unu ($\rho = 1$), atunci avem ceea ce se numește problema rădăcinii unitate, adică un model nestaționar, modelul „mers aleator". Ceea ce trebuie testat în acest caz este ipoteza $\rho = 1$.

Deoarece în soft-urile de statistică este mai dificil de testat ipoteza $\rho = 1$, o modalitate mai facilă de a testa staționaritatea pe baza problemei rădăcinii unitate este aceea a modelului care utilizează operatorul diferență de ordinul întâi:

$DY_t = Y_t - Y_{t-1}$.

Se consideră modelul echivalent:

$DY_t = (\rho - 1)Y_{t-1} + \varepsilon_t$ sau

$DY_t = \delta Y_{t-1} + \varepsilon_t$,

pentru care se va testa ipoteza $\delta = 0$.

Testarea ipotezei $\delta = 0$ se face cu ajutorul testului Dickey-Fuller. În esență, acest test este construit după modelul unei statistici de tip Student, însă valorile teoretice nu sunt date de distribuția Student, ci au fost generate de Dickey și Fuller prin simulare pe baza metodei Monte Carlo.

Ipotezele testului sunt:

$H_0: \delta = 0$;

$H_1: \delta < 0$.

În practică, testul Dickey-Fuller se aplică la unul din următoarele modele:

$DY_t = \delta Y_{t-1} + \varepsilon_t$, care este un model autoregresiv de ordinul întâi;

$DY_t = \beta_1 + \delta Y_{t-1} + \varepsilon_t$, care este un model autoregresiv cu constantă;

$DY_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + \varepsilon_t$, care este un model autoregresiv cu tendință.

Ultimul model cuprinde cele două tipuri de nestaționaritate: determinist și stocastic. Verificarea nestaționarității deterministe este o problemă relativ simplă, ea se face cu un test Student asupra coeficientului β. Verificarea nestaționarității aleatoare este mai dificilă și se realizează pe baza testului Dickey-Fuller.

[1] W.A. Fuller, *Introduction to Statistical Time Series*, John Wiley & Sons, New York, 1976.

Dacă este acceptată ipoteza $\delta = 0$, pentru fiecare model din cele de mai sus, procesul este nestaţionar.

Pentru cel de-al treilea model, decizia testului Dickey-Fuller se prezintă în tabelul de mai jos:

	$\delta < 0$	$\delta = 0$
$\beta_2 \neq 0$	staţionaritate stocastică nestaţionaritate deterministă	nestaţionaritate stocastică (rădăcină unitate) nestaţionaritate deterministă
$\beta_2 = 0$	staţionaritate stocastică staţionaritate deterministă	nestaţionaritate stocastică staţionaritate deterministă

Valorile critice ale testului Dickey-Fuller sunt diferite pentru cele trei modele de mai sus şi se citesc în funcţie de tipul de model utilizat.

Valoarea calculată a testului pentru parametrul δ se obţine după relaţia clasică a unui test Student, adică prin raportul dintre estimaţia parametrului şi estimaţia erorii standard:

$$t_{calc} = \frac{Estimation}{Std.Error}.$$

Observaţie

Dacă în modelele prezentate de Dickey şi Fuller, variabila reziduală admite fenomenul de autocorelare (ipoteză care se testează cu ajutorul testului Durbin-Watson), se construieşte un alt test, care poartă numele de testul Dickey-Fuller extins (*Augmented Dickey-Fuller Test - ADF Test*).

Acest test este dat prin relaţia:

$$DY_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + \alpha_i \sum_i DY_{t-i} + u_i.$$

Modelul de mai sus are rolul de a corecta autocorelarea erorilor din modelul Dickey-Fuller simplu. Operatorul DY_{t-i} ia valori pentru un număr de decalaje suficient de mare pentru a corecta autocorelarea erorilor ($DY_{t-1} = Y_{t-1} - Y_{t-2}$).

Exemplu

Se aplică testul Dickey-Fuller pentru seria de timp utilizată la exemplul anterior.

În R, testul admite implicit constanta şi componenta trend. Pentru componenta *augmented* se poate specifica un ordin anume sau acesta se identifică automat prin intermediul testului. Comanda şi rezultatele pentru variabila output sunt:

```
> adf.test(output)
```

```
Augmented Dickey-Fuller Test
data: output
Dickey-Fuller = -3,1366, Lag order = 6, p-value = 0,0989
alternative hypothesis: stationary
```

Rezultatul de mai sus ne prezintă valoarea calculată a testului și probabilitatea asociată. Deoarece *p-value > 0,05*, se decide să nu se respingă ipoteza nulă, adică seria analizată nu este staționară.

Testul Phillips-Perron (PP test)

Acesta este un alt test frecvent utilizat în practică pentru verificarea staționarității unei serii de timp. Valoarea calculată a testului se obține printr-o corecție a valorii calculate pentru testul DF. Pachetele program de statistică permit realizarea acestui test, ca și în cazul testului precedent.

Exemplu

Aplicăm testul PP la datele din exemplul anterior. Comanda și rezultatele obținute în R sunt:

```
> pp.test(output)

Phillips-Perron Unit Root Test
data: output
Dickey-Fuller Z(alpha) = -34,926, Truncation lag parameter = 5, p-value = 0,01
alternative hypothesis: stationary
```

Interesant este de observat că, pentru un risc de 5%, conform acestui test, se respinge ipoteza nulă sau ipoteza existenței unei rădăcini unitate (nestaționaritate).

5. Procese integrate

O serie de timp nestaționară poate fi transformată într-o serie staționară cu ajutorul operatorului diferență de un anumit ordin.

Definiție

Un proces stocastic nestaționar este integrat, dacă prin aplicarea operatorului diferență, se transformă într-un proces staționar.

Procesul se numește integrat de ordinul întâi, dacă prin operatorul diferență de ordinul întâi, se obține o serie staționară.

În general, pentru o serie care a fost diferențiată până la ordinul d și în urma diferențierii, se obține o serie staționară, spunem că avem o serie de timp integrată de ordin d. O asemenea serie se notează $Y_t \sim I(d)$.

Dacă în urma testării staționarității, după una dintre metodele prezentate anterior, se constată că seria de timp este nestaționară (rădăcină unitate), se verifică staționaritatea seriei obținute prin operatorul diferență de ordinul întâi. Procedura poate continua și în cazul în care seria nu este o serie integrată de ordinul întâi.

6. Procese cointegrate

Proprietatea de cointegrare se referă la seriile de timp integrate de un anumit ordin. De obicei, în practică, variabilele economice sunt integrate de ordinul 1. În acest caz, pentru două serii de timp I(1), se pune întrebarea ce proprietate are combinația liniară a acestor două variabile.

Definiție

Două procese stocastice se numesc cointegrate dacă sunt integrate de ordinul 1 și combinația lor liniară este staționară, adică este I(0).

De exemplu, dacă se consideră variabilele $Y_t \sim I(1)$ și $X_t \sim I(1)$, iar

$\varepsilon_t = Y_t - (\beta_0 + \beta_1 X_t) \sim I(0)$,

atunci cele două variabile sunt cointegrate.

Testul Engle-Granger

Acesta este un test de cointegrare[2] frecvent utilizat în practică. Ideea de bază a testului este aceea de a verifica staționaritatea variabilei reziduale rezultate din modelul de regresie dintre două variabile de timp nestaționare. Testul presupune așadar 2 etape:

- se verifică dacă cele 2 variabile analizate sunt integrate de ordinul 1;
- se testează ipoteza de staționaritate a variabilei reziduale din modelul:

$Y_t = \beta_0 + \beta_1 X_t + u_t$, adică se testează coeficientul δ din ecuația:

$Du_t = \delta u_{t-1} + \varepsilon_t$, unde

$u_t = \rho u_{t-1} + \varepsilon_t$ și

$\delta = (\rho - 1)$.

Dacă ρ este semnificativ (în valoare absolută mai mic decât 1), rezultă că erorile sunt staționare, adică variabilele sunt cointegrate.

Testul *tau* utilizat pentru a lua decizia este:

$$\tau = \frac{\hat{\rho} - 1}{\hat{\sigma}_{\hat{\rho}}}.$$

Exemplu

Utilizăm baza de date oferită de Yahoo, pentru perioada 2015-2016, privind prețurile lunare ale acțiunilor companiilor General Motors și Ford.

În R, se utilizează funcția *yegcm* care preia automat datele oferite de Yahoo. Comanda este următoarea:

```
< yegcm("F","GM", 20150101, 20161231)
```

[2] Engle, R.F., Granger C.W.J. (1987), Co-integration and Error Correction: Representation, Estimation and Testing, *Econometrica*, 55: 251-276

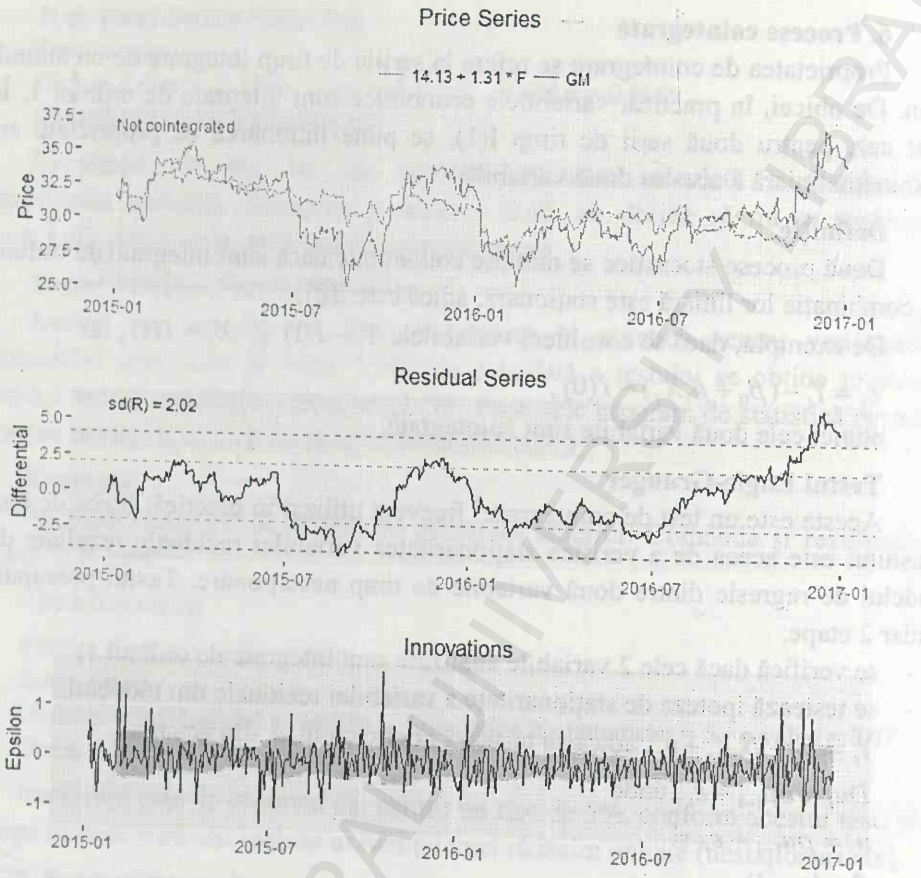


Figura 8.4. *Rezultatele analizei de cointegrare pentru prețurile Ford și General Motors*
Sursa: Realizată de autor pe baza datelor oferite de YAHOO, 2017

Rezultatele analizei sunt prezentate mai jos (în R, este menționat explicit rezultatul testului de cointegrare printr-o atenționare ce apare la final):

GM[i] = 1,3063 F[i] + 14,1281 + R[i], R[i] = 1,0000 R[i-1] + eps[i],
(0,0957) (1,2455) (0,0075)
eps ~ N(0, 0,3381^2)
R[2016-12-30] = 4,7654 (t = 2,362)
WARNING: F does not seem to be integrated. F and GM do not appear to be cointegrated.

R permite reprezentarea grafică a seriilor de date, împreună cu cele două variabile reziduale rezultate din modelare și testarea cointegrării, așa cum se observă din Figura 8.4 (graficele sunt construite prin aplicarea funcțiilor *print* și *plot* aplicate obiectului construit cu ajutorul funcției *yegcm*).

8.2. Modelarea seriilor de timp cu ajutorul proceselor stocastice

Scopul principal al modelării seriilor de timp cu ajutorul proceselor stocastice este de a explica modul de evoluţie a fenomenului şi de a realiza predicţii pe baza modelului estimat. Modelul funcţionează ca o schemă explicativă a fenomenului, având la dispoziţie evoluţia sa trecută, şi permite estimarea nivelului fenomenului pentru un orizont viitor.

În funcţie de tipul seriei de timp, în econometrie, se utilizează mai multe tipuri de procese stocastice ca modele pentru seriile de timp: procese autoregresive, procese de tip medie mobilă şi modele compuse pe baza acestora.

8.2.1. Procese de tip medie mobilă (MA)

Procesele de acest tip explică variaţia unei serii de timp cu ajutorul unei combinaţii liniare de procese gaussiene de tip *zgomot alb*. Prezentăm mai întâi procesul de ordinul 1 şi proprietăţile acestuia, şi apoi cazul general.

Proces MA(1)

Pentru o serie de timp Y_t, procesul de tip medie mobilă *de ordinul întâi* este dat de ecuaţia:

$$Y_t = \mu + \beta_0\varepsilon_t + \beta_1\varepsilon_{t-1}.$$

Proprietăţi

Dacă $\varepsilon_t \sim N(0,\sigma^2)$, atunci se obţin următorii parametri:

- media:
  $E(Y_t) = \mu$;
- varianţa:
  $\gamma_{0t} = E(Y_t - \mu)^2 = E(\beta_0\varepsilon_t + \beta_1\varepsilon_{t-1})^2 = (\beta_0^2 + \beta_1^2)\sigma^2$;
- funcţia de autocovarianţă de ordinul întâi:
  $\gamma_{1t} = E(Y_t - \mu)(Y_{t-1} - \mu) = E(\beta_0\varepsilon_t + \beta_1\varepsilon_{t-1})(\beta_0\varepsilon_{t-1} + \beta_1\varepsilon_{t-2}) = \beta_0 \cdot \beta_1\,\sigma^2$;
- funcţia de autocovarianţă de ordin k:
  $\gamma_{kt} = E(Y_t - \mu)(Y_{t-k} - \mu) = E(\beta_0\varepsilon_t + \beta_1\varepsilon_{t-1})(\beta_0\varepsilon_{t-k} + \beta_1\varepsilon_{t-k+1}) = 0$,
  pentru orice $k > 1$;
- funcţia de autocorelaţie de ordinul 1:
  $$\rho_1 = \frac{\beta_0\beta_1}{\beta_0^2 + \beta_1^2}.$$

Dacă $\beta_0 = 1$, $\rho_1 = \dfrac{\beta_1}{1+\beta_1^2}$;

- funcția de autocorelație de ordinul k:

$\rho_k = 0$, pentru orice $k > 1$.

Pe baza informațiilor de mai sus, se poate concluziona că procesul MA(1) este un proces staționar (vezi figura 8.5).

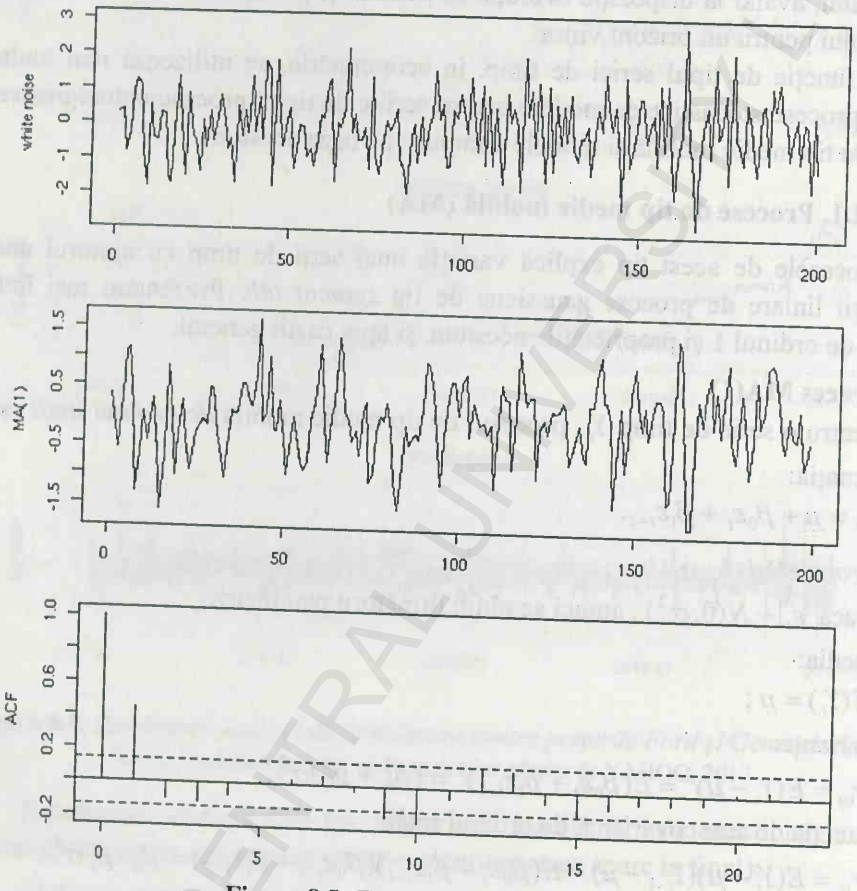


Figura 8.5. *Exemplu de proces MA(1)*
Sursa: Realizată de autor (date convenționale)

Proces MA(q)

Prin generalizarea procesului MA(1), se obține modelul MA(q):

$Y_t = \mu + \beta_0\varepsilon_t + \beta_1\varepsilon_{t-1} + ... + \beta_q\varepsilon_{t-q}$.

Proprietăți

$E(Y_t) = \mu$

$\gamma_{0t} = E(Y_t - \mu)^2 = E(\beta_0\varepsilon_t + \beta_1\varepsilon_{t-1} + ... + \beta_q\varepsilon_{t-q})^2 = (\beta_0^2 + \beta_1^2 + ... + \beta_q^2)\sigma^2$

BCU CLUJ CENTRAL UNIVERSITY LIBRARY

$$\gamma_{kt} = E(Y_t - \mu)(Y_{t-k} - \mu) = \begin{cases} 0, & k > q \\ (\beta_0\beta_k + \beta_1\beta_{k+1} + ... + \beta_{q-k}\beta_q)\sigma^2, & k = \overline{1,q} \end{cases}$$

$$\rho_k = \frac{\gamma_{kt}}{\gamma_{0t}} = \begin{cases} 0, & k > q \\ \dfrac{\beta_0\beta_k + \beta_1\beta_{k+1} + ... + \beta_{q-k}\beta_q}{\beta_0^2 + \beta_1^2 + ... + \beta_q^2}, & k = \overline{1,q} \end{cases}$$

Observaţie

Corelograma funcţiei de autocorelaţie a procesului MA(q) are primii q termeni diferiţi de zero. Pentru condiţia de staţionaritate, restul valorilor trebuie să fie în intervalul de încredere stabilit, adică nesemnificativ diferite de zero.

8.2.2. Procese autoregresive (AR)

Pentru procesele de acest tip, valoarea prezentă a variabilei de timp se obţine cu ajutorul unei combinaţii liniare de valori din trecut până la un anumit decalaj. Prezentăm mai întâi procesul de ordinul 1 şi proprietăţile acestuia, şi apoi cazul general.

Proces AR(1)

Se consideră o serie de timp Y_t. Seria poate fi modelată cu ajutorul procesului:

$$Y_t = \alpha_0 + \alpha\, Y_{t-1} + \varepsilon_t,$$

care este un model autoregresiv de ordinul întâi sau AR(1).

Modelul poate fi considerat şi fără constantă, adică:

$$Y_t = \alpha_1 Y_{t-1} + \varepsilon_t,$$

unde $\varepsilon_t \sim N(0,\sigma^2)$ este o variabilă de tip „zgomot alb", iar $E(Y_t) = \mu$.

Modelul fără constantă $Y_t = \alpha Y_{t-1} + \varepsilon_t$ se mai poate scrie în funcţie de operatorul decalaj:

$$(1 - \alpha L)Y_t = \varepsilon_t.$$

Altfel scris, procesul AR(1) devine:

$Y_t = \alpha Y_{t-1} + \varepsilon_t$ sau

$Y_t = \alpha^2 Y_{t-2} + \alpha\, \varepsilon_{t-1} + \varepsilon_t$ sau

$Y_t = \alpha^3 Y_{t-3} + \alpha^2 \varepsilon_{t-2} + \alpha\, \varepsilon_{t-1} + \varepsilon_t$ sau

$Y_t = \alpha^k Y_{t-k} + \alpha^{k-1} \varepsilon_{t-k-1} + ... + \alpha^2 \varepsilon_{t-2} + \alpha\, \varepsilon_{t-1} + \varepsilon_t$ sau

BCU CLUJ CENTRAL UNIVERSITY LIBRARY

$$Y_t = \sum_{i=0}^{\infty} \alpha^i \varepsilon_{t-i} = \sum_{i=0}^{\infty} \alpha^i L^{(i)} \varepsilon_t .$$

Această relație arată că efectul lui ε_k asupra lui y_t este:

$$\frac{\partial y_t}{\partial \varepsilon_k} = \alpha^k .$$

Acest efect poartă numele de *multiplicator dinamic*.
Relația este adevărată și pentru valori viitoare ale variabilei:

$$\frac{\partial y_{t+k}}{\partial \varepsilon_t} = \alpha^k .$$

Se poate observa că seria de mai sus este dezvoltarea în serie pentru:

$$\frac{1}{1-\alpha L} \text{ sau } (1-\alpha L)^{-1} .$$

Suma tuturor influențelor variabilelor reziduale sau *multiplicatorul total* este:

$$\sum_{i=0}^{\infty} \frac{\partial y_t}{\partial \varepsilon_i} = \sum_{i=0}^{\infty} \alpha^i = \frac{1}{1-\alpha}, \text{ dacă } |\alpha| < 1 .$$

Dacă utilizăm modelul cu constantă:
$Y_t = \alpha_0 + \alpha\, Y_{t-1} + \varepsilon_t$, modelul devine:

$$Y_t = (\alpha_0 + \varepsilon_t) + \alpha(\alpha_0 + \varepsilon_{t-1}) + \alpha^2(\alpha_0 + \varepsilon_{t-2}) + ... =$$
$$= \alpha_0(\alpha + \alpha^2 + ...) + \varepsilon_t + \alpha\varepsilon_{t-1} + \alpha^2\varepsilon_{t-2} + ...$$

Dacă $|\alpha| < 1$, rezultă:

$$Y_t = \alpha_0 \frac{1}{1-\alpha} + \varepsilon_t + \alpha\varepsilon_{t-1} + \alpha^2\varepsilon_{t-2} + ...$$

Proprietăți

$$E(Y_t) = \alpha_0 \frac{1}{1-\alpha};$$

$$V(Y_t) = V(\alpha_0 \frac{1}{1-\alpha} + \varepsilon_t + \alpha\varepsilon_{t-1} + \alpha^2\varepsilon_{t-2} + ...) =$$

$$= V(\varepsilon_t + \alpha\varepsilon_{t-1} + \alpha^2\varepsilon_{t-2} + ...) = \sigma^2(1 + \alpha^2 + \alpha^4 + ...) = \frac{\sigma^2}{1-\alpha^2},$$

pentru $|\alpha| < 1$;

BCU IAȘI CENTRAL UNIVERSITY LIBRARY

$$\gamma_{kt} = E(Y_t - \mu)(Y_{t-k} - \mu) = \frac{\alpha^k}{1-\alpha^2}\sigma^2;$$

$$\rho_k = \frac{\gamma_{0t}}{\gamma_{kt}} = \alpha^k.$$

Observaţie

Pentru procesul autoregresiv de ordinul întâi, funcţia de autocorelaţie descreşte exponenţial, iar condiţia de staţionaritate este asigurată dacă:

$|\alpha| < 1$.

Proces AR(p)

Prin generalizare, se obţine modelul autoregresiv de ordin *p*:

$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + \varepsilon_t$.

Un exemplu de proces autoregresiv de ordinul 3 este prezentat în figura 8.6.

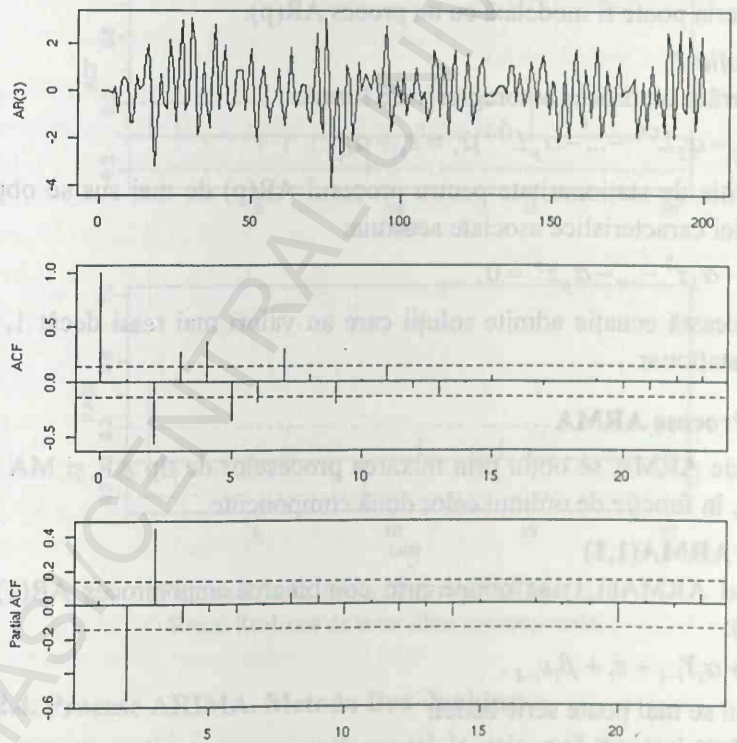


Figura 8.6. *Exemplu de proces AR(3)*
Sursa: Realizată de autor (date convenţionale)

Modelul se mai poate scrie:

BCU IAŞI CENTRAL UNIVERSITY LIBRARY

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + \varepsilon_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \varepsilon_t .$$

Dacă utilizăm operatorul decalaj, putem rescrie modelul:

$$(1 - \alpha_1 L - \alpha_2 L^{(2)} - ... - \alpha_p L^{(p)}) Y_t = \varepsilon_t + \alpha_0 .$$

Proprietăți

$$E(Y_t) = \frac{\alpha_0}{1 - \alpha_1 - \alpha_2 - ... - \alpha_p};$$

$$\rho_k = \alpha_1 \rho_{k-1} + \alpha_2 \rho_{k-2} + ... + \alpha_p \rho_{k-p} .$$

Observația 1

Procesul autoregresiv pentru o serie de timp poate fi identificat cu ajutorul corelogramei funcției de autocorelație parțială. Dacă în corelogramă, primii p termeni sunt diferiți de zero, iar restul sunt nesemnificativ diferiți de zero, se poate considera că seria poate fi modelată cu un proces AR(p).

Observația 2

Considerăm un proces autoregresiv de forma:

$$(1 - \alpha_1 L - \alpha_2 L^{(2)} - ... - \alpha_p L^{(p)}) Y_t = \varepsilon_t + \alpha_0 .$$

O condiție de staționaritate pentru procesul AR(p) de mai sus se obține cu ajutorul funcției caracteristice asociate acestuia:

$$1 - \alpha_1 z - \alpha_2 z^2 - ... - \alpha_p z^p = 0 .$$

Dacă această ecuație admite soluții care au valori mai mari decât 1, atunci procesul este staționar.

8.2.3. Procese ARMA

Procesele ARMA se obțin prin mixarea proceselor de tip AR și MA în mai multe moduri, în funcție de ordinul celor două componente.

Proces ARMA(1,1)

Procesul ARMA(1,1) se obține prin combinarea unui proces AR(1) și un proces MA(1):

$$Y_t = \theta + \alpha_1 Y_{t-1} + \varepsilon_t + \beta_1 \varepsilon_{t-1} .$$

Modelul se mai poate scrie astfel:

$(1 - \alpha_1 L) Y_t = (1 + \beta_1 L) \varepsilon_t$ sau

$$Y_t = \frac{(1 + \beta_1 L)}{(1 - \alpha_1 L)} \varepsilon_t .$$

BCU "ION CREANGĂ" UNIVERSITY LIBRARY

Observaţii

- procesul ARMA(1,1) este staţionar pentru $|\alpha_1| < 1, |\beta_1| < 1$ (vezi figura 8.7);
- funcţia de autocorelaţie descreşte exponenţial, începând cu $k = 1$, în timp ce pentru un proces AR(1), aceasta descreşte exponenţial, începând cu $k = 0$.

Proces ARMA(p,q)

Prin generalizare, se obţine un model ARMA(p,q):

$Y_t = \theta + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + \beta\varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q}$ sau

$(1 - \alpha_1 L - \alpha_2 L^{(2)} - ... - \alpha_p L^{(p)}) Y_t = \theta + (1 + \beta_1 L + \beta_2 L^{(2)} + ... + \beta_q L^{(q)}) \varepsilon_t$.

Observaţie

Staţionaritatea procesului ARMA(p,q) depinde de parametrii autoregresivi: $\alpha_1, \alpha_2, ..., \alpha_p$.

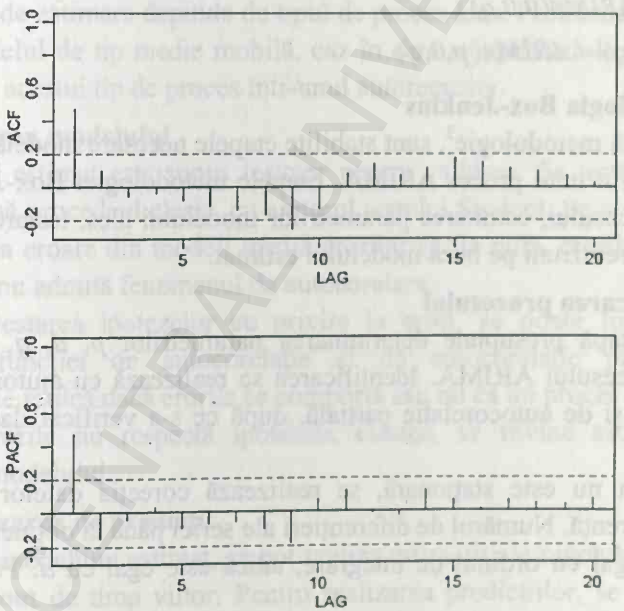


Figura 8.7. *Exemplu de proces ARMA(1,1)*
Sursa: Realizată de autor (date convenţionale)

8.2.4. Procese ARIMA. Metoda Box-Jenkins

În această secţiune, se prezintă modelele cele mai generale care combină procese de tip medie mobilă şi procese autoregresive, ţinând cont de proprietatea de staţionaritate a seriilor de timp analizate.

1. Procese ARIMA (p,d,q)

Procesele ARIMA sunt construite ca o generalizare a proceselor ARMA, pentru seriile care nu respectă condiția de staționaritate.

O serie de timp nestaționară stocastic se poate staționariza prin aplicarea operatorului diferență. Dacă seria este diferențiată până la un ordin d pentru a fi staționarizată, aceasta se numește serie integrată de ordin d.

Procesul ARIMA este determinat de trei parametri:

- ordinul componentei autoregresive p;
- ordinul de integrare d;
- ordinul componentei medie mobilă q.

Procesul ARIMA(p,d,q) este modelul cel mai general al unei serii de timp. Celelalte modele sunt cazuri particulare ale acestui tip de proces. Adică:

$AR(p) = ARIMA(p,0,0)$;

$MA(q) = ARIMA(0,0,q)$;

$ARMA(p,q) = ARIMA(p,0,q)$.

2. Metodologia Box-Jenkins

Prin această metodologie[3], sunt stabilite etapele necesare modelării unei serii de timp cu ajutorul unui proces ARIMA. Etapele metodologiei Box-Jenkins sunt: identificarea procesului, estimarea parametrilor modelului ales, testarea modelului și realizarea de previziuni pe baza modelului estimat.

a. Identificarea procesului

Această etapă presupune determinarea parametrilor p, d, q, care permit identificarea procesului ARIMA. Identificarea se realizează cu ajutorul funcțiilor de autocorelație și de autocorelație parțială, după ce s-a verificat dacă seria este staționară.

Dacă seria nu este staționară, se realizează corecția datelor cu ajutorul operatorului diferență. Numărul de diferențieri ale seriei până la obținerea unei serii staționare este egal cu ordinul de integrare, adică este egal cu d. Pentru o serie staționară, $d = 0$.

Pentru seria staționară (reală sau integrată), se construiesc cele două corelograme, iar analiza acestora permite identificarea tipului de proces care este cel mai adecvat seriei de date.

Există următoarele situații posibile:

- pentru o corelogramă a funcției de autocorelație care descrește exponențial, iar corelograma funcției de autocorelație parțială are primii p termeni

[3] G.E. Box, G.M. Jenkins, *Time series analysis: forcasting and control*, Holden Day, 1976.

semnificativ diferiți de zero, în timp ce restul sunt în intervalul de încredere, se identifică un model AR(p);

- pentru o corelogramă care are primii q termeni semnificativ diferiți de zero pentru funcția de autocorelație, restul fiind nesemnificativi statistic, iar corelograma funcției de autocorelație parțială descrește exponențial, se identifică un model MA(q);
- modelul ARMA se propune atunci când nu este foarte evident un tip de proces din cele două de mai sus sau când corelogramele indică un proces care implică ambele componente (ambele diagrame au un număr de coeficienți semnificativi).

b. Estimarea parametrilor modelului

După ce s-a stabilit tipul de proces pentru seria de date disponibilă, se trece la etapa de estimare a parametrilor modelului.

Metoda de estimare depinde de tipul de proces ales. Probleme practice ridică îndeosebi modelul de tip medie mobilă, caz în care se apelează la proprietatea de transformare a acestui tip de proces într-unul autoregresiv.

c. Testarea modelului

Modelul estimat este supus testelor pentru validare. Se testează parametrii modelului, după procedeul clasic, cu ajutorul testului Student. Se verifică ipotezele pentru variabila eroare din model: media erorilor să fie nulă, erorile să fie normal repartizate, să nu admită fenomenul de autocorelare.

Pentru testarea ipotezelor cu privire la erori, se poate folosi cu succes corelograma funcției de autocorelație și de autocorelație parțială. Forma corelogramei ne indică dacă erorile se comportă sau nu ca un proces aleator pur.

Dacă erorile nu respectă ipotezele clasice, se revine asupra etapei de identificare a modelului.

d. Realizarea de predicții

Pe baza modelului estimat, se pot realiza estimații ale nivelului fenomenului pentru un orizont de timp viitor. Pentru realizarea predicțiilor, se ține seama de toate transformările pe care le-a suportat seria de timp în procesul de modelare.

Prognoza se poate realiza punctual sau se pot obține intervale de încredere sau zone de prognoză pentru orizontul stabilit.

Exemplu

Considerăm seria de timp formată din înregistrările oferite de Yahoo, pentru prețul săptămânal al acțiunilor companiei IBM, între 7 aprilie 2014 și 10 aprilie 2017. Seria este reprezentată grafic în figura 8.8.

Așa cum o arată diagrama din figura 8.8, este posibil ca seria să fie nestaționară stocastic. Aceeași concluzie se poate desprinde analizând diagrama funcției de autocorelație din figura 8.9. În acest caz, se aplică demersul metodologiei Box-Jenkins pentru a estima un model econometric și pentru a realiza o predicție cu privire la următoarele săptămâni din aprilie 2017.

Identificare

Analiza tipului de nestaționaritate se realizează pe baza testului Dickey-Fuller.

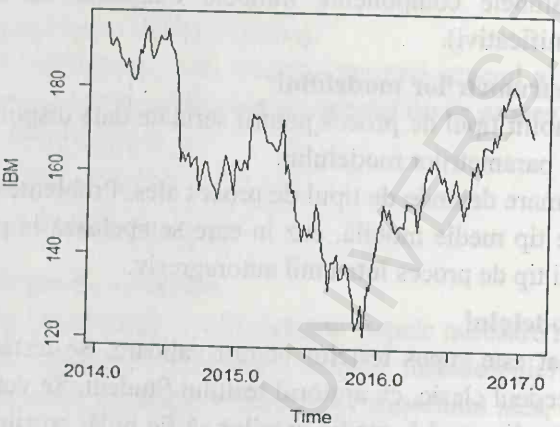


Figura 8.8. *Tendința de evoluție a prețului săptămânal al acțiunilor IBM, între aprilie 2014 și aprilie 2017*

Sursa: Realizată de autor pe baza datelor oferite de YAHOO, 2017

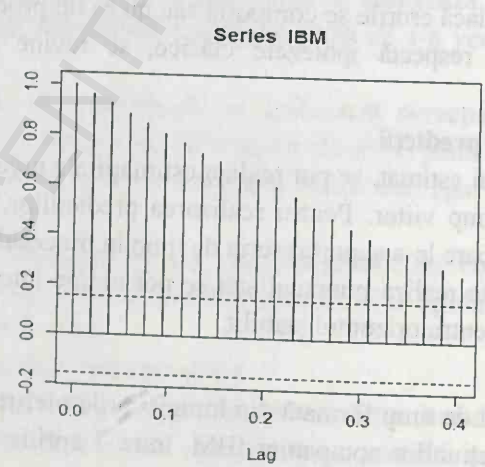


Figura 8.9. *Corelograma funcției de autocorelație*

Sursa: Realizată de autor pe baza datelor oferite de YAHOO, 2017

BCU IAȘI / CENTRUL UNIVERSITAR LIBRARY

Rezultatele testului Dickey-Fuller sunt prezentate mai jos.

```
> adf.test(IBM)

Augmented Dickey-Fuller Test
data:  IBM
Dickey-Fuller = -1,4901,    Lag order = 5,  p-value = 0,789
alternative hypothesis: stationary
```

Pentru un prag de semnificație de 0,05, nu se respinge ipoteza nulă $\rho = 1$, adică se poate spune că seria analizată este nestaționară.

Rezultatele testului Dickey-Fuller sugerează aplicarea operatorului diferență de ordinul întâi pentru seria de date analizată și reluarea testului pentru noua serie. În R, se obține:

```
> adf.test(diff(IBM,1))

Augmented Dickey-Fuller Test
data:  diff(IBM, 1)
Dickey-Fuller = -4,8369,    Lag order = 5,  p-value = 0,01
alternative hypothesis: stationary
```

Concluzia este că seria transformată cu ajutorul operatorului de ordinul 1 este staționară, adică seria inițială este integrată de ordinul 1. Corelograma funcției de autocorelație și a celei de autocorelație parțială pentru seria diferențiată arată același lucru (vezi figura 8.10):

```
> par(mfrow=c(2,1))
> acf(diff(IBM,1))
> pacf(diff(IBM,1))
```

Ținând cont de diferențierea seriei din etapa precedentă, pentru modelarea seriei de timp inițiale, se pot utiliza mai multe variante de procese: ARIMA(2,1,1), ARIMA(1,1,1) sau ARIMA(1,1,0). Alegem să exemplificăm ultimul caz.

Pentru un model cu constantă, s-au obținut rezultatele de mai jos.

```
> arima(IBM, order=c(1,1,0))

Call:
arima(x = IBM, order = c(1, 1, 0))
Coefficients:
ar1..-0,1092
s.e.   0,0796
```

Modelul estimat este:

$DY_t = -0{,}109 DY_{t-1}$.

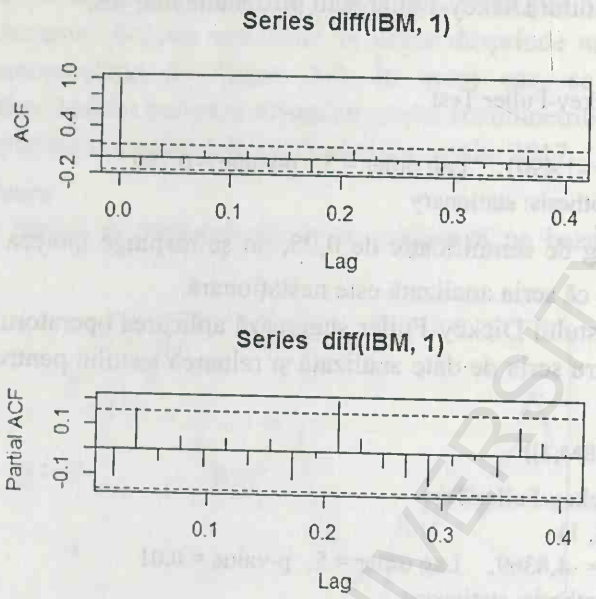


Figura 8.10. *Corelogramele seriei diferențiate*
Sursa: Realizată de autor pe baza datelor oferite de YAHOO, 2017

Validare

Pentru validarea modelului ARIMA(1,1,0), se verifică staționaritatea erorilor acestui model.

```
> adf.test(res)

Augmented Dickey-Fuller Test
data: res
Dickey-Fuller = -4, 7822,  Lag order = 5,  p-value = 0,01
alternative hypothesis: stationary
```

Conform rezultatelor testului ADF, se poate accepta ipoteza de staționaritate a erorilor.

Predicție

Pe baza modelului estimat, se pot obține valori prognozate pentru următoarele 2 săptămâni.

```
> predict(arima(IBM, order=c(1,1,0)), n.ahead=2)
$pred

Time Series:
Start = c(2017, 5)
```

BCU "ALEXANDRU IOAN CUZA" UNIVERSITY LIBRARY

End = c(2017, 6)
Frequency = 52
[1] 169,8150; 169,7839
$se
Time Series:
Start = c(2017, 5)
End = c(2017, 6)
Frequency = 52
[1] 4,024463; 5,389689

Cu ajutorul modelului ARIMA(1,1,0), s-au estimat prețurile medii ale acțiunilor IBM pentru următoarele 2 săptămâni: 169,8150, respectiv 169,7839 de dolari.

8.3. Teste și exerciții

Teste grilă

1. Un proces este staționar dacă:
a) media, varianța și funcția de autocorelație sunt constante în timp;
b) media și varianța sunt constante în timp;
c) corelogramele funcțiilor de autocorelație și autocorelație parțială au coeficienți nesemnificativi statistic.

2. Un *zgomot alb* este un proces:
a) staționar;
b) nestaționar;
c) cu dispersie zero;
d) cu medie zero.

3. Care din următoarele modele sunt staționare?
a) $Y_t = \mu + \varepsilon_t + \beta_1 \varepsilon_{t-1}$;
b) $Y_t = \mu + Y_t + \varepsilon_t$;
c) $Y_t = 2{,}5 + 0{,}7Y_{t-1}$;
d) $Y_t = 1{,}2 - 1{,}5\varepsilon_{t-1}$.

4. Care din următoarele modele sunt nestaționare?
a) $Y_t = \mu + Y_t + \varepsilon_t$;
b) $Y_t = 2{,}1 - 0{,}5Y_{t-1}$;
c) $Y_t = \mu + \varepsilon_t + \beta_1 \varepsilon_{t-1}$;

BCU ASE CENTRAL UNIVERSITY LIBRARY

d) $Y_t = Y_{t-1} + \varepsilon_t$.

5. Pentru modelul de regresie $Y_t = 2{,}1 - 0{,}5Y_{t-1}$, sunt valabile afirmațiile:

a) seria este autoregresivă de ordinul 1;

b) seria este de tip medie mobilă de ordinul 1;

c) corelograma funcției de autocorelație parțială are un coeficient semnificativ;

d) corelograma funcției de autocorelație are un coeficient semnificativ.

6. Un proces de tip *zgomot alb* are valorile funcției de autocorelație estimate în intervalul:

a) $(-1{,}96; 1{,}96)$;

b) $(-1{,}96\frac{1}{\sqrt{n}}; 1{,}96\frac{1}{\sqrt{n}})$;

c) $(-1{,}96\sqrt{n}; 1{,}96\sqrt{n})$;

d) $(-1{,}96\frac{\sigma}{\sqrt{n}}; 1{,}96\frac{\sigma}{\sqrt{n}})$.

7. Pentru o serie de timp, s-au obținut rezutatele:

Augmented Dickey-Fuller Test
data: output
Dickey-Fuller = -2,14 Lag order = 6, p-value = 0,129
alternative hypothesis: stationary

Au loc enunțurile:

a) seria de timp este staționară;

b) valoarea testului DF este în regiunea de acceptare;

c) seria de timp este nestaționară;

d) seria de timp este integrată.

8. Un *proces ARIMA* este:

a) staționar;

b) nestaționar;

c) o combinație de două procese staționare.

9. Proprietatea de cointegrare presupune:

a) posibilitatea construirii unui model econometric pentru două variabile nestaționare integrate de ordinul 1;

b) o combinație de două procese staționare;

BCU IAȘI / CENTRAL UNIVERSITY LIBRARY

c) existența unei combinații liniare staționare dintre două serii integrate de ordinul 1.

10. Testarea staționarității unei serii se poate realiza:
a) grafic, pe baza funcțiilor de autocorelație și autocorelație parțială;
b) numeric, cu ajutorul testelor Phillips-Perron și Dickey-Fuller;
c) grafic, cu ajutorul corelogramei liniare.

Exerciții

1. Prezentați, pe scurt, diferența dintre seriile de timp staționare și cele nestaționare. Se poate observa diferența și pe cale grafică? Argumentați.

2. Similar, prezentați diferența dintre un proces de tip MA și unul de tip AR. Cum evaluăm această diferență pe cale grafică?

3. Se dă următorul model estimat: $Y_t = -0,5Y_{t-1} + 0,25Y_{t-2}$. Utilizând condiția de staționaritate pentru un proces autoregresiv (funcția caracteristică), să se determine dacă seria pentru care s-a obținut modelul este staționară sau nu.

4. Se consideră seria de date, oferită de Yahoo, privind prețul acțiunilor companiei Ford în intervalul 10 iunie 2010 - 10 iunie 2017. Se cere:
- să se reprezinte grafic seria de timp utilizând R;
- să se reprezinte grafic cele două corelograme, pentru funcțiile de autocorelație și autocorelație parțială, pentru cel puțin 24 de lag-uri. Comentați rezultatul.
- modelați evoluția seriei de timp utilizând metodologia Box-Jenkins;
- estimați prețul mediu al acțiunilor pentru următoarele două luni.

5. Pentru exemplul din secțiunea dedicată metodologiei Box-Jenkins, au rămas spre analiză alte două modele posibile: ARIMA(2,1,1), ARIMA(1,1,1).
Construiți aceste modele și comparați rezultatele. Cum se alege cel mai bun model dintre cele 3 care pot explica evoluția în timp a prețului acțiunilor IBM?

BCU IASI CENTRALA UNIVERSITY LIBRARY

Capitolul 9

MODELE ECONOMETRICE DINAMICE

În modelarea econometrică, se utilizează o categorie de modele de regresie care se construiesc pentru a studia relații de cauzalitate ce apar în timp. Rezultă că variabile ce se utilizează în analiză sunt variabile de timp. Relațiile dinamice dintre fenomene presupun că dependența dintre variabile, la un moment dat, este determinată și de valorile variabilelor la momente de timp anterioare. Modelele de acest tip se numesc *modele econometrice dinamice*.

Există două categorii de modele dinamice:

- *Modele autoregresive*. În aceste modele, valoarea variabilei dependente la un moment dat depinde de valorile pe care această variabilă le ia la momente anterioare de timp.
- *Modele cu decalaj distribuit*. În aceste modele, valoarea variabilei dependente la un moment dat depinde de valorile variabilei independente de la același moment și de la momente de timp anterioare.

În acest capitol, se prezintă câteva aspecte legate de specificul și rolul modelelor dinamice în economie, precum și câteva probleme legate de estimarea și testarea acestor modele.

9.1. Specificul și importanța modelelor dinamice

Importanța studiului relațiilor de cauzalitate dinamică în economie poate fi ușor demonstrată prin apelul la câteva exemple: impactul variațiilor de temperatură asupra recoltelor agricole și a prețurilor produselor agricole; influența creșterilor salariale pe o perioadă de timp asupra consumului și economiilor unei familii; relația dintre dinamica investițiilor și creșterea economică etc.

BCU ASE LIBRARY

Modele econometrice dinamice care pot surprinde asemenea relații de cauzalitate în timp sunt prezentate mai jos.

1. Prezentarea modelelor

Modelele din această categorie depind de orizontul de timp din trecut care este luat în considerare în relațiile dintre variabile.

Modele autoregresive

În funcție de numărul de valori anterioare de care depinde variabila dependentă, modelele autoregresive pot fi:

- modele de ordinul 1, date prin relația:
  $Y_t = \alpha_0 + \beta X_t + \gamma Y_{t-1} + \varepsilon_t$;
- modele de ordinul k:
  $Y_t = \alpha_0 + \beta X_t + \gamma_1 Y_{t-1} + \gamma_2 Y_{t-2} + ... + \gamma_k Y_{t-k} + \varepsilon_t$.

Modele cu decalaj distribuit

Similar, în funcție de numărul de valori ale variabilei independente din trecut, modelele cu decalaj distribuit sunt:

- modele cu decalaj distribuit de ordinul 1:
  $Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$;
- modele cu decalaj distribuit de ordinul k:
  $Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_k X_{t-k} + \varepsilon_t$.

2. Decalajul în economie

În economie, legătura dintre variabila dependentă (Y) și variabila independentă (X), de obicei, nu este instantanee. Între acțiunea variabilei independente și efectul său – măsurat prin variabila dependentă – există, de regulă, o perioadă de timp, un decalaj (în limba engleză, este utilizat termenul *lag*).

Putem identifica mai multe motive ale apariției decalajului pentru fenomenele economice:

- temeiuri politice – schimbările politice dint-o anumită țară se pot reflecta imediat în dinamica piețelor financiare;
- motive psihologice – inerția unui comportament sau a unui mod de viață, temeri, suspiciuni ale actorilor economici;
- temeiuri tehnologice – de exemplu, evoluția tehnologiei pe piața IT sau a tehnologiilor din industria chimică;
- motive instituționale – legi, contracte, acorduri internaționale etc.

Pentru construirea modelelor econometrice, este important să se cunoască modul în care operează asemenea mecanisme, să se identifice factorii care pot

BCU CENTRALA UNIVERSITY LIBRARY

acționa într-o anumită situație și apoi să se specifice variabilele care pot intra în componența modelelor.

9.2. Estimarea modelelor dinamice

În acest subcapitol, se vor prezenta problemele specifice estimării pentru fiecare tip de model econometric dinamic în parte.

9.2.1. Estimarea modelelor cu decalaj distribuit

Se consideră modelul de regresie cu decalaj distribuit de ordin k:

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_k X_{t-k} + \varepsilon_t .$$

Variabilele $X_{t-i}, i = \overline{0,k}$, se numesc variabile exogene și respectă condiția:

$$E(\varepsilon_t / X_{t-i}) = 0, i = \overline{0,k} .$$

Parametrii $\beta_i, i = \overline{0,k}$, semnifică influența variabilei independente pentru fiecare moment de timp, din prezent până la un moment k din trecut, iar α_0 este media variabilei dependente atunci când influența variabilei independente este nulă.

Parametrul β_0 măsoară efectul contemporan sau imediat al variabilei independente asupra celei dependente:

$$\beta_0 = \frac{\partial Y_t}{\partial X_t} \text{ sau } \beta_0 = \partial Y_t \Big|_{\partial X_t = 1} .$$

Similar, parametrii $\beta_i, i = \overline{1,k}$, măsoară impactul valorilor variabilei din trecut, la diverse momente, asupra valorii prezente a variabilei dependente:

$$\beta_i = \frac{\partial Y_t}{\partial X_{t-i}} .$$

Multiplicatori

Pentru acest model, se construiesc sumele parțiale ale parametrilor $\beta_i, i = \overline{1,k}$, pentru perioade de timp diferite:

β_0,

$\beta_0 + \beta_1$,

...

$\beta_0 + \beta_1 + ... + \beta_k$.

Aceste sume reprezintă modificările care se realizează (în medie) asupra variabilei dependente pentru o perioadă de timp anume (t, $t-1$, $t-2$ etc.).

Coeficienții β_i poartă numele de *multiplicatori dinamici* și se diferențiază astfel:

- β_0 este multiplicatorul de impact sau contemporan și este denumit *short-run multiplier*;
- $\beta_i, i=\overline{1,k}$, sunt multiplicatori de moment;
- sumele parțiale ne dau multiplicatorii *intermediari* (*interim multipliers*);
- suma multiplicatorilor după k perioade, $\beta = \sum_{i=0}^{k} \beta_i$, se numește *multiplicator total* sau *cumulat* (*long-run multiplier*), adică impactul variabilei independente asupra celei dependente pe termen lung.

Fiecare multiplicator poate fi standardizat și se obține relația:

$$\beta_i^* = \frac{\beta_i}{\sum_{i=0}^{k} \beta_i}.$$

Acest indicator arată cât din influența lui X asupra lui Y este determinată de influența din perioada „i".

Identificarea ordinului decalajului

Pentru modelele cu decalaj distribuit, este important să se cunoască numărul de decalaje sau până la ce nivel în timp există o influență a valorilor din trecut asupra variabilei dependente. Teoretic, numărul de decalaje poate fi finit sau infinit, deși, în practică, influența se poate studia până la un anumit ordin.

Pentru a identifica tipul de model sau ordinul de decalaj, se poate utiliza o metodă secvențială, care presupune construirea de modele succesive, prin introducerea unei noi variabile cu decalaj superior. Procedeul se poate opri atunci când o nouă variabilă introdusă în model este nesemnificativă.

Metode de estimare a modelelor cu decalaj distribuit

Estimarea modelelor cu decalaj distribuit trebuie să țină cont de o problemă delicată care apare, de obicei, la aceste modele: coliniaritatea variabilelor independente.

Pentru estimarea parametrilor, se utilizează mai multe metode, în funcție de numărul de decalaje. Pentru modele cu un decalaj care tinde la infinit, se utilizează frecvent metoda Koyck[1], iar pentru modelele cu un decalaj finit, se aplică metoda Almon[2].

[1] L.M. Koyck, *Distributed lags and investment analysis*, North-Holland, Amsterdam, 1954.

[2] S. Almon, „The distributed lag between capital appropriation and expenditures", *Econometrica*, nr. 1, 1965.

În general, aceste metode presupun, în final, utilizarea metodei celor mai mici pătrate de estimare a parametrilor. Aplicarea acestei metode presupune respectarea ipotezelor clasice ale unui model de regresie, aşa cum au fost prezentate în capitolul 7: variabilele independente să fie necoliniare; erorile să nu fie corelate cu variabilele independente la momentul *t*; erorile să aibă media zero, să fie homoscedastice, să nu fie corelate pentru diferite momente de timp, să respecte legea de distribuţie normală.

1. Metoda Koyck

Metoda Koyck presupune că între multiplicatori există o relaţie de tipul unei progresii geometrice:

$\beta_k = \beta_0 \lambda^k$, pentru $k = 0, 1, 2...$ şi $\lambda \in (0,1)$, unde

λ este rata de declin sau de decalaj;

$1 - \lambda$ este viteza de ajustare.

Ipoteza lui Koyck se traduce astfel:

- deoarece $\lambda > 0$, coeficienţii de regresie vor avea acelaşi semn;
- deoarece $\lambda < 1$, cu cât se introduce în model un nou lag, cu atât efectul său asupra variabilei dependente va fi mai mic;
- chiar dacă numărul de lag-uri este infinit, multiplicatorul total este finit, adică:

$$\beta = \sum_{k=0}^{\infty} \beta_k = \beta_0 \sum_{k=0}^{\infty} \lambda^k = \beta_0 \frac{1}{1-\lambda}.$$

Se consideră modelul cu decalaj de ordin *k*:

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_k X_{t-k} + \varepsilon_t .$$

Prin transformarea Koyck, se obţine:

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_0 \lambda X_{t-1} + \beta_0 \lambda^2 X_{t-2} + ... + \beta_0 \lambda^k X_{t-k} + \varepsilon_t .$$

Pentru modelul anterior, se realizează o transformare cu ajutorul operatorului decalaj sau lag de ordinul 1:

$$Y_{t-1} = \alpha_0 + \beta_0 X_{t-1} + \beta_0 \lambda X_{t-2} + \beta_0 \lambda^2 X_{t-3} + ... + \beta_0 \lambda^k X_{t-k+1} + \varepsilon_t .$$

Înmulţind ultimul model cu λ şi scăzând din cel anterior, se obţine:

$$Y_t - \lambda Y_{t-1} = \alpha_0 (1-\lambda) + \beta_0 X_t + \varepsilon_t - \lambda \varepsilon_{t-1} .$$

Notând $u_t = \varepsilon_t - \lambda \varepsilon_{t-1}$, rezultă modelul:

$$Y_t = \alpha_0 (1-\lambda) + \beta_0 X_t + \lambda Y_{t-1} + u_t ,$$

care este un model autoregresiv de ordinul 1.

Aplicarea metodei lui Koyck conduce la următoarele probleme:

- obținerea unui model autoregresiv, ceea ce înseamnă că, printre variabilele independente, apare variabila stocastică Y_{t-1};
- corelarea variabilelor Y_{t-1} și u_t;
- autocorelarea erorilor $\varepsilon_t : u_t = \varepsilon_t - \lambda\varepsilon_{t-1}$.

Estimarea modelului cu decalaj distribuit după metoda Koyck se rezumă, în final, la a estima un model de tip autoregresiv, subiect ce va fi analizat într-o secțiune ulterioară.

2. Metoda Almon

Se consideră modelul de regresie cu decalaj de ordinul k:

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_k X_{t-k} + \varepsilon_t = \alpha_0 + \sum_{i=0}^{k} \beta_i X_{t-i} + \varepsilon_t .$$

Metoda Almon are la bază ipoteza că un multiplicator se poate explica printr-un polinom de forma:

$$\beta_i = a_0 + a_1 i + a_2 i^2 + ... + a_p i^p .$$

De regulă, se utilizează transformarea pentru $p = 2$:

$\beta_i = a_0 + a_1 i + a_2 i^2$, adică:

$\beta_0 = a_0$,

$\beta_1 = a_0 + a_1 + a_2$,

$\beta_2 = a_0 + 2a_1 + 4a_2$,

...

$\beta_k = a_0 + ka_1 + k^2 a_2$.

Prin înlocuire în modelul inițial, se obține modelul:

$$Y_t = \alpha_0 + a_0 \sum_{i=0}^{k} X_{t-i} + a_1 \sum_{i=0}^{k} iX_{t-i} + a_2 \sum_{i=0}^{k} i^2 X_{t-i} + \varepsilon_t .$$

Notând:

$$Z_{0t} = \sum_{i=0}^{k} X_{t-i} ,$$

$$Z_{1t} = \sum_{i=0}^{k} iX_{t-i} ,$$

$$Z_{2t} = \sum_{i=0}^{k} i^2 X_{t-i} ,$$

se obține modelul:

BCU CLUJ CENTRAL UNIVERSITY LIBRARY

$$Y_t = \alpha_0 + a_0 Z_{0t} + a_1 Z_{1t} + a_2 Z_{2t} + \varepsilon_t,$$

care este un model de regresie liniară multiplă.

Aplicând MCMMP, se obţin estimatorii: $\hat{a}_0, \hat{a}_1, \hat{a}_2$, iar pe baza lor, se determină estimatorii multiplicatorilor:

$$\hat{\beta}_0 = \hat{a}_0,$$

$$\hat{\beta}_1 = \hat{a}_0 + \hat{a}_1 + \hat{a}_2,$$

$$\hat{\beta}_2 = \hat{a}_0 + 2\hat{a}_1 + 4\hat{a}_2,$$

...

$$\hat{\beta}_k = \hat{a}_0 + k\hat{a}_1 + k^2\hat{a}_2.$$

Problemele pe care le ridică metoda Almon sunt:

- alegerea numărului de lag-uri *k*;
- alegerea gradului din transformarea $\beta_i = a_0 + a_1 i + a_2 i^2 + ... + a_p i^p$.

Ca soluţie, în practică, se determină un număr mediu de lag-uri, prin diferite metode, iar $p = 2$.

Exemplu

Utilizăm baza de date oferită de Fondul Monetar Internaţional, pentru perioada 1990-2012, privind indicele producţiei industriale (*output*, *Y*) şi indicele preţurilor de consum (*inflation*, *X*). Baza de raportare este anul 2005, iar frecvenţa datelor este lunară.

Seriile de timp se creează în R cu ajutorul funcţiei *ts()*. În sintaxa acesteia se adaugă: variabila numerică (*vector*), momentul de start, prin precizarea anului şi lunii, momentul de final şi frecvenţa seriei. Pentru exemplul utilizat, rezultă comanda:

```
> out=ts(output, start=c(1990,10), end=c(2012,7), frequency=12)
```

Cele 2 serii sunt reprezentate grafic în figura 9.1 cu ajutorul funcţiei *plot*.

Pentru variabilele considerate, se construieşte un model cu decalaj distribuit de forma:

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \varepsilon_t.$$

Prin metoda Almon, se construiesc variabilele:

$$Z_{0t} = \sum_{i=0}^{3} X_{t-i} = X_t + X_{t-1} + X_{t-2} + X_{t-3},$$

$$Z_{1t} = \sum_{i=0}^{3} iX_{t-i} = X_{t-1} + 2X_{t-2} + 3X_{t-3},$$

BCU IASI CENTRAL UNIVERSITY LIBRARY

$$Z_{2t} = \sum_{i=0}^{3} i^2 X_{t-i} = X_{t-1} + 4X_{t-2} + 9X_{t-3}.$$

Variabilele de mai sus se construiesc în R cu ajutorul funcției decalaj: *lag(ts,k)*.

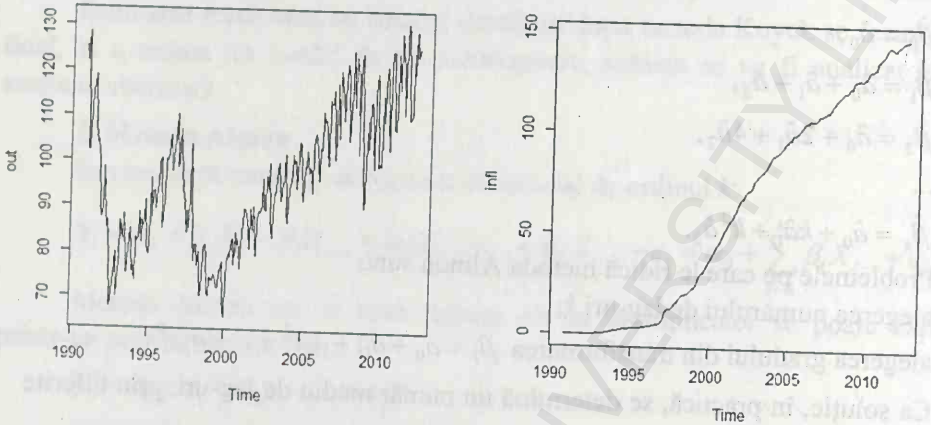


Figura 9.1. *Reprezentarea grafică a variabilelor utilizate output și inflation*
Sursa: Realizată de autor pe baza datelor oferite de FMI, 2014

Estimațiile parametrilor modelului $Y_t = \alpha_0 + a_0 Z_{0t} + a_1 Z_{1t} + a_2 Z_{2t} + \varepsilon_t$ se prezintă în tabelul de mai jos.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	88,360	1,119	78,931	<2e-16 ***
Z0	-1,712	1,128	-1,518	0,130
Z1	-2,810	3,155	-0,890	0,374
Z2	1,710	1,048	1,632	0,104

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
Residual standard error: 9.876 on 255 degrees of freedom
Multiple R-squared: 0,5978, Adjusted R-squared: 0,593
F-statistic: 126,3 on 3 and 255 DF, p-value: < 2,2e-16

Pe baza estimațiilor a_0, a_1, a_2, se calculează estimațiile modelului inițial astfel:

$b_0 = a_0 = -1,71$;

$b_1 = a_0 + a_1 + a_2 = -1,71 - 2,81 + 1,71 = -2,81$;

$b_2 = a_0 + 2a_1 + 4a_2 = -1,71 - 5,62 + 6,84 = -0,49$;

$b_3 = a_0 + 3a_1 + 9a_2 = -1,71 - 8,43 + 15,39 = 5,25$.

BCU IASI CENTRAL UNIVERSITY LIBRARY

În final, modelul estimat prin metoda Almon este:

$Y_t = 88{,}36 - 1{,}71X_t - 2{,}81X_{t-1} - 0{,}49X_{t-2} + 5{,}25X_{t-3}$.

9.2.2. Estimarea modelelor autoregresive

Prin metoda Koyck, se obţine un model autoregresiv, iar estimarea acestuia presupune rezolvarea problemelor enunţate anterior, probleme care ţin de încălcarea ipotezelor modelului clasic de regresie.

Modelul autoregresiv de ordinul 1 poate fi scris astfel:

$Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 Y_{t-1} + u_t$.

În estimarea modelului autoregresiv, apar o serie de probleme, dintre care cele mai importante sunt: corelarea variabilei eroare cu variabila dependentă de un anumit decalaj şi autocorelarea erorilor.

a. Problema corelării variabilelor Y_{t-1} şi u_t

O soluţie la această problemă poate fi utilizarea variabilelor instrumentale. Această metodă presupune ca, în modelul de mai sus, în locul variabilei Y_{t-1} să fie utilizată, de exemplu, variabila X_{t-1}.

Practic, estimarea parametrilor din modelul $Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 Y_{t-1} + u_t$ se realizează înlocuind variabila Y_{t-1} cu variabila instrumentală doar în a treia ecuaţie din sistemul de ecuaţii normale obţinut prin aplicarea MCMMP.

Aşadar, parametrii se obţin schimbând ultima ecuaţie din sistemul:

$$\begin{cases} n\hat{\alpha}_0 + \hat{\alpha}_1 \sum_t X_t + \hat{\alpha}_2 \sum_t Y_{t-1} = \sum_t Y_t \\ \hat{\alpha}_0 \sum_t X_t + \hat{\alpha}_1 \sum_t X_t^2 + \hat{\alpha}_2 \sum_t X_t Y_{t-1} = \sum_t X_t Y_t \\ \hat{\alpha}_0 \sum_t Y_{t-1} + \hat{\alpha}_1 \sum_t X_t Y_{t-1} + \hat{\alpha}_2 \sum_t Y_{t-1}^2 = \sum_t Y_{t-1} Y_t \end{cases}$$

Prin transformare, se obţine sistemul:

$$\begin{cases} n\hat{\alpha}_0 + \hat{\alpha}_1 \sum_t X_t + \hat{\alpha}_2 \sum_t Y_{t-1} = \sum_t Y_t \\ \hat{\alpha}_0 \sum_t X_t + \hat{\alpha}_1 \sum_t X_t^2 + \hat{\alpha}_2 \sum_t X_t Y_{t-1} = \sum_t X_t Y_t \\ \hat{\alpha}_0 \sum_t X_{t-1} + \hat{\alpha}_1 \sum_t X_t X_{t-1} + \hat{\alpha}_2 \sum_t X_{t-1} Y_{t-1} = \sum_t X_{t-1} Y_t \end{cases}$$

Rezolvând sistemul de mai sus, obţinem $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2$, estimatori nedeplasaţi şi convergenţi (însă posibil să fie neeficienţi) pentru parametrii modelului.

b. Problema autocorelării erorilor în modelul autoregresiv

Se consideră modelul autoregresiv:

$Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 Y_{t-1} + u_t$.

Detectarea autocorelației erorilor se face cu ajutorul testului *h Durbin*.

Etapele testării sunt:

1. Formularea ipotezelor

H_0: nu există autocorelație;

H_1: există autocorelație.

2. Se alege pragul de semnificație

3. Se estimează parametrii modelului, fără a ține seama de autocorelație, și se determină o estimație pentru $V(\hat{\alpha}_2)$.

4. Se utilizează statistica

$$h = (1 - \frac{d}{2}) \sqrt{\frac{n}{1 - n\hat{V}(\hat{\alpha}_2)}},$$

unde d este statistica Durbin-Watson, pentru care $\hat{\rho} = 1 - \frac{d}{2}$.

În condițiile acceptării ipotezei nule, statistica h are o lege de distribuție asimptotic normală.

5. Decizia se ia în funcție de valoarea calculată a statisticii h. De exemplu, dacă $h_{calc} \in [-1,96; 1,96]$, atunci nu se respinge ipoteza H_0.

Dacă modelul nu înregistrează fenomenul de autocorelare a erorilor, estimarea parametrilor se realizează clasic, prin metoda celor mai mici pătrate.

Estimarea modelului autoregresiv în cazul existenței autocorelării erorilor

În practică, se utilizează mai multe metode: metoda de corecție iterativă a lui Cochrane-Orcutt, metoda modelului de quasi-diferență etc.

a. Metoda iterativă

Se consideră modelul autoregresiv:

$Y_t = \alpha_0 + \beta X_t + \gamma Y_{t-1} + \varepsilon_t$, în care

$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$, $|\rho| < 1$.

Procedeul de estimare presupune parcurgerea metodei Cochrane - Orcutt. După testarea existenței autocorelării erorilor cu ajutorul testului *h Durbin*, se aplică procedeul iterativ de estimare, utilizând estimațiile coeficientului de corelație a erorilor cu ajutorul estimatorului:

BCU CLUJ UNIVERSITY LIBRARY

$$\hat{\rho} = 1 - \frac{d}{2}.$$

b. Metoda modelului de quasi-diferenţă

O altă cale facilă este utilizarea metodei modelului de quasi-diferenţă. Această metodă presupune parcurgerea următoarelor etape:

- Construirea modelului de quasi-diferenţă:

  $Y_t = \alpha_0 + \beta X_t + \gamma Y_{t-1} + \varepsilon_t$,

  $Y_{t-1} = \alpha_0 + \beta X_{t-1} + \gamma Y_{t-2} + \varepsilon_{t-1}$.

  Înmulţind a doua ecuaţie cu ρ şi scăzând din prima, rezultă modelul:

  $Y_t - \rho Y_{t-1} = \alpha_0(1-\rho) + \beta(X_t - \rho X_{t-1}) + \gamma(Y_{t-1} - \rho Y_{t-2}) + \varepsilon_t - \rho\varepsilon_{t-1}$ sau

  $Y_t - \rho Y_{t-1} = \alpha_0(1-\rho) + \beta(X_t - \rho X_{t-1}) + \gamma(Y_{t-1} - \rho Y_{t-2}) + u_t$ sau

  $Y_t = \alpha_0(1-\rho) + \beta(X_t - \rho X_{t-1}) + (\gamma+\rho)Y_{t-1} - \rho Y_{t-2} + u_t$.

- Construirea estimatorului pentru parametrul ρ.

  Scris altfel, ultimul model este:

  $Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 Y_{t-1} + \beta_4 Y_{t-2} + u_t$, unde

  $\beta_0 = \alpha_0(1-\rho)$, $\beta_1 = \beta$, $\beta_2 = -\beta\rho$, $\beta_3 = (\gamma+\rho)$, $\beta_4 = -\rho$.

  În consecinţă,

  $\rho = -\frac{\beta_2}{\beta_1}$, iar estimatorul său va fi:

  $$\hat{\rho} = -\frac{\hat{\beta}_2}{\hat{\beta}_1}.$$

- Estimarea modelului de quasi-diferenţă cu ajutorul estimatorului construit mai sus:

  $Y_t - \hat{\rho} Y_{t-1} = \alpha_0(1-\hat{\rho}) + \beta(X_t - \hat{\rho} X_{t-1}) + \gamma(Y_{t-1} - \hat{\rho} Y_{t-2}) + u_t$ sau

  $Y_t^* = \alpha^* + \beta X_t^* + \gamma Y_{t-1}^* + u_t$, unde

  $Y_t^* = Y_t - \hat{\rho} Y_{t-1}$,

  $\alpha^* = \alpha_0(1-\hat{\rho})$,

  $X_t^* = X_t - \hat{\rho} X_{t-1}$.

Ultimul model oferă posibilitatea estimării parametrilor modelului autoregresiv în condiţii de calitate şi respectând ipotezele modelului clasic de regresie.

Exemplu

Utilizăm datele de la exemplul din secţiunea anterioară.

Se estimează modelul autoregresiv $Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 Y_{t-1} + u_t$ și se verifică ipoteza de autocorelare a erorilor.

În tabelul de mai jos, este calculată estimația abaterii standard a estimatorului parametrului $\alpha_2 : s_{\hat{\alpha}_2} = 0{,}034$.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16,67	3,033	5,49	0,000
inf	0,045	0,010	4,30	0,000
lag(out, 1)	0,798	0,034	22,82	0,000

Testul Durbin - Watson pentru variabila reziduală este prezentat mai jos:

```
> dwt(res)
```

[1] 2,163787

Pe baza estimațiilor de mai sus, se calculează estimația statisticii *h-Durbin* cu ajutorul relației:

$$h_{calc} = (1 - \frac{d}{2}) \sqrt{\frac{n}{1 - n s^2_{\hat{\alpha}_2}}}$$. Rezultă:

$$h_{calc} = (1 - \frac{2{,}16}{2}) \sqrt{\frac{259}{1 - 259 \cdot 0{,}034^2}} = -1{,}53$$

Deoarece $h_{calc} > -1{,}96$, pentru modelul autoregresiv propus, nu există fenomenul de autocorelare a erorilor, iar modelul nu trebuie corectat.

9.3. Teste și exerciții

Teste grilă

1. În modelul de regresie $Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$:

a) β_0 este multiplicatorul de impact;

b) β_0 este ordonata la origine;

a) α_0 este multiplicatorul de impact;

b) α_0 este ordonata la origine.

2. Cauzalitatea dinamică în economie este determinată de:

a) natura variabilelor economice;

b) factorii conjuncturali;

c) variațiile unor factori aleatori ce nu pot fi estimați;

d) acțiunea determinanților politici și legislativi.

BCU "MIHAIL KOGĂLNICEANU" CENTRAL UNIVERSITY LIBRARY

3. Modelele cu decalaj distribuit:
a) au la bază o combinație liniară a valorilor prezente și trecute ale variabilei independente;
b) admit problema coliniarității variabilelor independente;
c) se bazează pe valori prezente și trecute ale variabilei dependente.

4. Într-un model cu decalaj distribuit, multiplicatorul total arată:
a) impactul variabilei independente asupra celei dependente pe termen scurt;
c) media variabilei dependente când variabilele independente sunt nule;
b) impactul variabilei independente asupra celei dependente pe termen lung.

5. Estimarea modelelor autoregresive ridică problema:
a) corelării variabilelor Y_{t-1} și ε_t;
c) autocorelării variabilelor ε_t;
b) independenței variabilelor ε_t.

Exerciții

1. Prezentați, pe scurt, diferența dintre următoarele modele econometrice:
$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$;
$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$;
$Y_t = \alpha_0 + \beta X_t + \gamma Y_{t-1} + \varepsilon_t$.

2. Cu ajutorul metodei Almon, pentru două variabile s-a estimat modelul de regresie:
$Y_t = 12{,}25 + 1{,}14 X_t + 0{,}54 X_{t-1} + 0{,}12 X_{t-2} - 0{,}43 X_{t-3}$.
Se cere:
- să se interpreteze estimațiile parametrilor;
- să se calculeze și interpreteze multiplicatorul total;
- metoda Almon permite calculul erorilor standard pentru parametrii din modelul $Y_t = \alpha_0 + a_0 Z_{0t} + a_1 Z_{1t} + a_2 Z_{2t} + \varepsilon_t$.
Cum se pot calcula erorile standard pentru estimațiile de mai sus care corespund parametrilor modelului inițial:
$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \varepsilon_t$?

3. Pentru un model de regresie de forma $Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 Y_{t-1} + u_t$, s-au obținut estimațiile de mai jos la nivelul unui eșantion de $n = 90$ de înregistrări:

Coefficients:

	Estimate	Std. Error
(Intercept)	-10,20	3,033
X_t	0,35	0,09
Y_{t-1}	0,18	0,06
DW test	1,70	

Folosind statistica *h*, se cere să se verifice dacă pentru modelul estimat există sau nu fenomenul de autocorelare a erorilor.

BCU IASI/CENTRAL UNIVERSITY LIBRARY

Bibliografie

Almon, S. (1965). The distributed lag between capital appropriations and expenditures, *Econometrica*, 33(1), pp. 178-196

Andrei, T. (2003). *Statistică şi econometrie*, Editura Economică, Bucureşti

Andrei, T., Bourbonnais, R. (2008). *Econometrie*, Editura Economică, Bucureşti

Andren, T. (2007). *Econometrics*, Ventus Publishing, Telluride, Colorado

Anderson, O.D. (1976). *Time Series Analysis and Forecasting*, Butterworth & Co Publishers Ltd, London

Arellano, M. (2003). *Panel Data Econometrics*, Oxford University Press, Oxford

Baltagi, B.H. (2005). *Econometric Analysis of Panel Data*, 3rd Edition, John Wiley & Sons, England

Barreto, H., Howland, F. (2005). *Introductory Econometrics*, Cambridge University Press, Cambridge

Bera, A.K., Jarque, C.M. (1981). An Efficient Large-Sample Test for Normality of Observations and Regression Residuals, Working Papers in *Econometrics*, 40, 31 p., Australian National University, Canberra

Berdot, J.P. (2001). *Économétrie*, CNED, Futurscope, Poitiers

Bender, E.A. (1978). *An Introduction to Mathematical Modeling*, John Wiley & Sons, New York

Blaug, M. (1992). *Teoria economică în retrospectivă*, Editura Didactică şi Pedagogică, Bucureşti

Bourbonnais, R. (2000). *Économétrie*, Dunod, Paris

Box, G.E., Jenkins, G.M. (1970). *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco

Box, G.E., Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*, Revised Edition, Holden Day, San Francisco

Breusch,T.S., Pagan A.R. (1979), A Simple Test for Heteroskedasticity and Random Coefficient Variation, *Econometrica*, 47(5), pp. 1287-1294

BCU CLUJ UNIVERSITY LIBRARY

Brooks, C. (2008). *Introductory Econometrics for Finance*, Cambridge University Press, Cambridge

Chirilă, V. (2013). *Analiza econometrică a burselor de valori și a ciclurilor de afaceri*, Editura ASE, București

Chow, G. (1989). *Econometrics*, MacGraw-Hill, New York

Dalgaard, P. (2008). *Introductory Statistics with R*, Springer, New York

Daudin, J., Robin, S. (1999). *Statistique inférentielle*, Presses Universitaires de Rennes, France

Dickey, D.A., Fuller W.A. (1979). Distributions of the Estimators for Autoregressive Time Series with a Unit Root, *Journal of the American Statistical Association*, 74(366), pp. 427-431

Dobb, J.L. (1953). *Stochastic Process*, John Wiley & Sons, New York

Dougherty, C. (2007). *Introduction to Econometrics*, Oxford University Press, Oxford

Eatwell, J., Newman, P. (1996). *Times Series and Statistics*, W.W. Norton, New York

Enders, W. (2015). *Applied Econometric Time Series*, 4th Edition, John Wiley & Sons, New York

Engle, R.F., Granger C.W.J. (1987), Co-integration and Error Correction: Representation, Estimation and Testing, *Econometrica*, 55(2), pp. 251-276

Field, A., Miles, J., Field, Z. (2012). *Discovering Statistics Using R*, Sage Publications Ltd, London

Fisher, R.A. (1947). *Les méthodes statistiques*, Presses Universitaires de France, Paris

Fuller, W.A. (1976). *Introduction to Statistical Time Series*, John Wiley & Sons, New York

Galton, F. (1886). Family Likeness in Stature, *Proceedings of Royal Society*, London, 40(242-245), pp. 42-73

Georgescu-Roegen, N. (1998). *Metoda statistică*, Editura Expert, București

Gourieroux, C., Monfort, A. (1989). *Statistique et modèles économétriques*, Economica, Paris, 2 vol.

Granger, C.W.J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods, *Econometrica*, 37(3), pp. 424-438

Granger, C., (1981). Some Properties of Time Series Data and Their Use in Econometric Model Specification, *Journal of Econometrics*, 16(1), pp. 121-130

Greene, W.H. (2003). *Econometric Analysis*, Pearson Education Publisher, New Jersey

Goldfeld, S.M., Quandt, R.E. (1972). *Non Linear Methods in Econometrics*, North-Holland, Amsterdam

Gujarati, D.N. (2004). *Basic Econometrics*, McGraw-Hill, New York

Hamilton, J.D. (1994). *Time Series Analysis*, Princeton University Press, New Jersey

Hendry, D.F. (2000). *Econometrics. Alchemy or Science*, Oxford University Press, Oxford

Hausman, D.M., ed. (1993). *Filosofia ştiinţei economice*, Antologie, Editura Humanitas, Bucureşti

Ionescu, H.M. (1962). *Introducere în statistica matematică*, Editura Didactică şi Pedagogică, Bucureşti

Isaic-Maniu, Al., Mitruţ, C., Voineagu,V. (2004). *Statistica*, Editura Universitară, Bucureşti

Iacob, A.I., Tanasoiu, O. (2005). *Modele econometrice*, Editura ASE, Bucureşti

Jaba, E., *Statistica* (2002). Ediţia a treia, Editura Economică, Bucureşti

Jaba, E., Jemna, D.V. (2006). *Econometrie*, Editura Sedcom Libris, Iaşi

Jemna, D.V. (2012). *Econometrie*, Editura Sedcom Libris, Iaşi

Jemna, D.V. (2015). *Sondajul statistic*, Editura Sedcom Libris, Iaşi

Johnston, J. (1963). *Econometric Methods*, MacGraw-Hill, New York

Johnston, J., Dinardo, J. (1997). *Méthodes économétriques*, Economica, Paris

Kendall, M.G., Stuart, A. (1963). *The Advanced Theory of Statistics*, Charles Griffin, London, 3 vol.

Kenedy, P. (1998). *A Guide to Econometrics*, 4th Edition, MIT Press, Cambridge

Keynes, J.M. (1936). *The General Theory of Employment, Interest and Money*, Macmillan Publishing Company, London

Keuzenkamp, H.A. (2004). *Probability, Econometrics and Truth*, Cambridge University Press, Cambridge

Kleiber, C., Zeileis, A. (2008). *Applied Econometrics with R*, Springer, New York

Kmenta, J. (1986). *Elements of Econometrics*, 2nd Edition, Macmillan Publishing Company, New York

Koyck, L.M. (1954). *Distributed Lags and Investment Analysis*, North-Holland, Amsterdam

Kozhan, R. (2009). *Financial Econometrics - with Eviews*, Ventus Publishing, Denmark

Labrose, C. (1976). *Introduction à l'économétrie*, Dunod, Paris

Laplace, P.S. (1812), *Théorie analytique des probabilités*, Ve. Courcier, Paris

Le Gall, Ph. (2007). *A History of Econometrics in France: From nature to models*, Routledge, London and New York

Maddala, G.S. (1978). *Econometrics*, MacGraw-Hill Book Company, New York
Maddala, G.S. (2001). *Introduction to Econometrics*, 3rd Edition, John Wiley & Sons, New York
Malinvaud, E. (1978). *Méthodes statistiques de l'économétrie*, Dunod, Paris
McEachern, W.A. (2000). *Economics: A Contemporary Introduction*, South-Western College Publishing, Mason, Ohio, USA
Mihoc, G., Craiu, V. (1976). *Tratat de statistică matematică*, Volum I, Editura Academiei Române, București
Monfort, A. (1997). *Cours de statistique mathématique*, Economica, Paris
Nenciu, E. (1984). *Teoria probabilităților și statistică matematică*, Editura Universității „Al.I. Cuza" din Iași
Onicescu, O. (1971). *Principes de logique et de philosophie mathématique*, București, Editura Academiei Române
Pecican, E.S. (1994). *Econometria*, Editura All, București
Pecican, E.S. (2003). *Econometria pentru economiști*, Editura Economică, București
Pintilescu, C. (2007). *Analiză statistică multivariată*, Editura Universității „Al.I. Cuza" din Iași
Popper, K.R. (1981). *Logica cercetării*, Editura Științifică și Enciclopedică, București
Samuelson, P., Koopmans, T., Stone, R. (1954). Report of the Evaluative Committee for Econometrica, *Econometrica*, 22(2), pp. 141-146
Sheather, S. (2009). *A Modern Approach to Regression with R*, Springer, New York
Stock, J., Watson, M. (2011). *Introduction to Econometrics*, 3rd Edition, Pearson Education Limited, Essex, England
Stigler, S.M. (1990), *The History of Statistics*, Harvard University Press, Cambridge
Stowell, S. (2014). *Using R for Statistics*, Apress, Berkeley, California
Studenmund, A.H. (2006). *Using Econometrics. A Practical Guide*, Pearson Education Limited, Essex, England
Tassi, P. (1989). *Méthodes statistiques*, Economica, Paris
Tijms, H. (2003). *A First Course in Stochastic Models*, John Wiley & Sons, Chichester, United Kingdom
Turturean, C.I. (2006). *Metode statistice de analiză a seriilor de timp*, Editura Sedcom Libris, Iași
Țigănescu, E., Roman, M. (2005). *Macroeconomie: o abordare cantitativă*, Editura Economica, București

BCU CENTRALA UNIVERSITARA LIBRARY

Wang, P. (2009). *Financial Econometrics*, Routledge, New York

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica*, 48(4), pp. 817-838

Wilthien, P.H. (1996). *Décision statistique et économétrie*, Armand Colin, Paris

Wooldridge, J.M. (2000). *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, London

Wooldridge, J.M. (2000). *Introductory Econometrics*, South-Western College Publishing, Mason, Ohio, USA

Yule, U.G., Kendall, M.C. (1969). *Introducere în teoria statisticii*, Editura Ştiinţifică, Bucureşti

Zaiţ, D., Nica, P. (1995). *Introducere în modelarea econometrică*, Editura Universităţii „Al.I. Cuza" din Iaşi

BCU IASI/CENTRAL UNIVERSITY LIBRARY

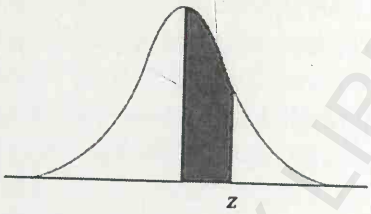
BCU IASI/CENTRAL UNIVERSITY LIBRARY

Tabele probabiliste

BCU IASI/CENTRAL UNIVERSITY LIBRARY

Funcția Laplace

$$\Phi(z)=\int_0^z e^{\frac{t^2}{2}}dt$$

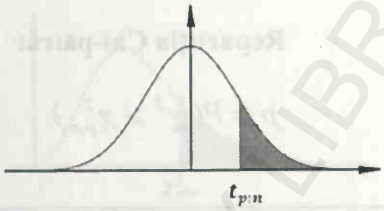


	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

BCU LIBRARY

Repartiţia Student

$$p = P(t > t_{p,n})$$

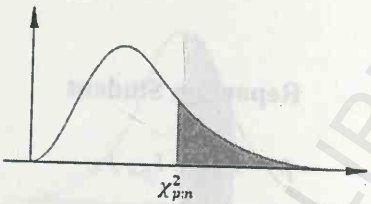


n\p	0,10	0,05	0,025	0,01	0,005
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
n>30	1,282	1,645	1,960	2,326	2,576

BCU IASI / CENTRAL UNIVERSITY LIBRARY

Repartiția Chi-pătrat

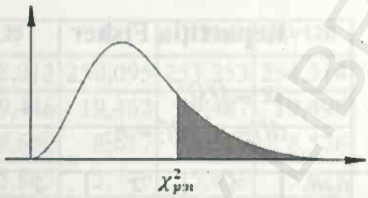
$$p = P(\chi^2 > \chi^2_{p,n})$$



n\p	,100	,050	,025	,010	,005
1	2,706	3,841	5,024	6,635	7,879
2	4,605	5,991	7,378	9,210	10,597
3	6,251	7,815	9,348	11,345	12,838
4	7,779	9,488	11,143	13,277	14,860
5	9,236	11,071	12,833	15,086	16,750
6	10,645	12,592	14,449	16,812	18,548
7	12,017	14,067	16,013	18,475	20,278
8	13,362	15,507	17,535	20,090	21,955
9	14,684	16,919	19,023	21,666	23,589
10	15,987	18,307	20,483	23,209	25,188
11	17,275	19,675	21,920	24,725	26,757
12	18,549	21,026	23,337	26,217	28,300
13	19,812	22,362	24,736	27,688	29,819
14	21,064	23,685	26,119	29,141	31,319
15	22,307	24,996	27,488	30,578	32,801
16	23,542	26,296	28,845	32,000	34,267
17	24,769	27,587	30,191	33,409	35,718
18	25,989	28,869	31,526	34,805	37,156
19	27,204	30,144	32,852	36,191	38,582
20	28,412	31,410	34,170	37,566	39,997
21	29,615	32,671	35,479	38,932	41,401
22	30,813	33,924	36,781	40,289	42,796
23	32,007	35,172	38,076	41,638	44,181
24	33,196	36,415	39,364	42,980	45,559
25	34,382	37,652	40,646	44,314	46,928
26	35,563	38,885	41,923	45,642	48,290
27	36,741	40,113	43,195	46,963	49,645
28	37,916	41,337	44,461	48,278	50,993
29	39,087	42,557	45,722	49,588	52,336
30	40,256	43,773	46,979	50,892	53,672

Repartiția Chi-pătrat

$$p = P(\chi^2 > \chi^2_{p,n})$$



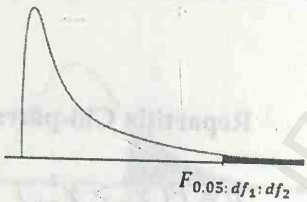
n\p	,995	,975	,950	,990	,900
1	0,00004	0,00016	0,00098	0,00393	0,01579
2	0,01003	0,02010	0,05064	0,10259	0,21072
3	0,07172	0,11483	0,21580	0,35185	0,58437
4	0,20699	0,29711	0,48442	0,71072	1,06362
5	0,41174	0,55430	0,83121	1,14548	1,61031
6	0,67573	0,87209	1,23734	1,63538	2,20413
7	0,98926	1,23904	1,68987	2,16735	2,83311
8	1,34441	1,64650	2,17973	2,73264	3,48954
9	1,73493	2,08790	2,70039	3,32511	4,16816
10	2,15586	2,55821	3,24697	3,94030	4,86518
11	2,60322	3,05348	3,81575	4,57481	5,57778
12	3,07382	3,57057	4,40379	5,22603	6,30380
13	3,56503	4,10692	5,00875	5,89186	7,04150
14	4,07467	4,66043	5,62873	6,57063	7,78953
15	4,60092	5,22935	6,26214	7,26094	8,54676
16	5,14221	5,81221	6,90766	7,96165	9,31224
17	5,69722	6,40776	7,56419	8,67176	10,08519
18	6,26480	7,01491	8,23075	9,39046	10,86494
19	6,84397	7,63273	8,90652	10,11701	11,65091
20	7,43384	8,26040	9,59078	10,85081	12,44261
21	8,03365	8,89720	10,28290	11,59131	13,23960
22	8,64272	9,54249	10,98232	12,33801	14,04149
23	9,26042	10,19572	11,68855	13,09051	14,84796
24	9,88623	10,85636	12,40115	13,84843	15,65868
25	10,51965	11,52398	13,11972	14,61141	16,47341
26	11,16024	12,19815	13,84390	15,37916	17,29188
27	11,80759	12,87850	14,57338	16,15140	18,11390
28	12,46134	13,56471	15,30786	16,92788	18,93924
29	13,12115	14,25645	16,04707	17,70837	19,76774
30	13,78672	14,95346	16,79077	18,49266	20,59923

BCU IAȘI DIGITAL LIBRARY

Repartiția Fisher

$\alpha = 0,05$

$df_1 = n_1$, $df_2 = n_2$



n_2/n_1	1	2	3	4	5	6	7
1	161,448	199,500	215,707	224,583	230,162	233,986	236,768
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353
3	10,128	9,552	9,277	9,117	9,014	8,941	8,887
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094
5	6,608	5,786	5,410	5,192	5,050	4,950	4,876
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787
8	5,318	4,459	4,066	3,838	3,688	3,581	3,501
9	5,117	4,257	3,863	3,633	3,482	3,374	3,293
10	4,965	4,103	3,708	3,478	3,326	3,217	3,136
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764
15	4,543	3,682	3,287	3,056	2,901	2,791	2,707
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514
21	4,325	3,467	3,073	2,840	2,685	2,573	2,488
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388
27	4,210	3,354	2,960	2,728	2,572	2,459	2,373
28	4,196	3,340	2,947	2,714	2,558	2,445	2,359
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334
40	4,085	3,232	2,839	2,606	2,450	2,336	2,249
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087
n>120	3,842	2,996	2,605	2,372	2,214	2,099	2,010

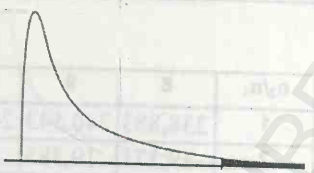
BCU LIBRARY

n_2/n_1	8	9	10	20	30	120	$n_1>120$
1	238,883	240,543	241,882	248,013	250,095	253,253	254,314
2	19,371	19,385	19,396	19,446	19,462	19,487	19,496
3	8,845	8,812	8,786	8,660	8,617	8,549	8,526
4	6,041	5,999	5,964	5,803	5,746	5,658	5,628
5	4,818	4,773	4,735	4,558	4,496	4,399	4,365
6	4,147	4,099	4,060	3,874	3,808	3,705	3,669
7	3,726	3,677	3,637	3,445	3,376	3,267	3,230
8	3,438	3,388	3,347	3,150	3,079	2,967	2,928
9	3,230	3,179	3,137	2,937	2,864	2,748	2,707
10	3,072	3,020	2,978	2,774	2,700	2,580	2,538
11	2,948	2,896	2,854	2,646	2,571	2,448	2,405
12	2,849	2,796	2,753	2,544	2,466	2,341	2,296
13	2,767	2,714	2,671	2,459	2,380	2,252	2,206
14	2,699	2,646	2,602	2,388	2,308	2,178	2,131
15	2,641	2,588	2,544	2,328	2,247	2,114	2,066
16	2,591	2,538	2,494	2,276	2,194	2,059	2,010
17	2,548	2,494	2,450	2,230	2,148	2,011	1,960
18	2,510	2,456	2,412	2,191	2,107	1,968	1,917
19	2,477	2,423	2,378	2,156	2,071	1,930	1,878
20	2,447	2,393	2,348	2,124	2,039	1,896	1,843
21	2,421	2,366	2,321	2,096	2,010	1,866	1,812
22	2,397	2,342	2,297	2,071	1,984	1,838	1,783
23	2,375	2,320	2,275	2,048	1,961	1,813	1,757
24	2,355	2,300	2,255	2,027	1,939	1,790	1,733
25	2,337	2,282	2,237	2,008	1,919	1,768	1,711
26	2,321	2,266	2,220	1,990	1,901	1,749	1,691
27	2,305	2,250	2,204	1,974	1,884	1,731	1,672
28	2,291	2,236	2,190	1,959	1,869	1,714	1,654
29	2,278	2,223	2,177	1,945	1,854	1,698	1,638
30	2,266	2,211	2,165	1,932	1,841	1,684	1,622
40	2,180	2,124	2,077	1,839	1,744	1,577	1,509
60	2,097	2,040	1,993	1,748	1,649	1,467	1,389
120	2,016	1,959	1,911	1,659	1,554	1,352	1,254
$n_2>120$	1,938	1,880	1,831	1,571	1,459	1,221	1,000

Repartiția Fisher

$\alpha = 0,01$

$df_1 = n_1$, $df_2 = n_2$



n_2/n_1	1	2	3	4	5	6	7
1	4052,181	4999,500	5403,352	5624,583	5763,650	5858,986	5928,356
2	98,503	99,000	99,166	99,249	99,299	99,333	99,356
3	34,116	30,817	29,457	28,710	28,237	27,911	27,672
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178
9	10,561	8,022	6,992	6,422	6,057	5,802	5,613
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026
17	8,400	6,112	5,185	4,669	4,336	4,102	3,927
18	8,285	6,013	5,092	4,579	4,248	4,015	3,841
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699
21	8,017	5,780	4,874	4,369	4,042	3,812	3,640
22	7,945	5,719	4,817	4,313	3,988	3,758	3,587
23	7,881	5,664	4,765	4,264	3,939	3,710	3,539
24	7,823	5,614	4,718	4,218	3,895	3,667	3,496
25	7,770	5,568	4,675	4,177	3,855	3,627	3,457
26	7,721	5,526	4,637	4,140	3,818	3,591	3,421
27	7,677	5,488	4,601	4,106	3,785	3,558	3,388
28	7,636	5,453	4,568	4,074	3,754	3,528	3,358
29	7,598	5,420	4,538	4,045	3,725	3,499	3,330
30	7,562	5,390	4,510	4,018	3,699	3,473	3,304
40	7,314	5,179	4,313	3,828	3,514	3,291	3,124
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953
120	6,851	4,787	3,949	3,480	3,174	2,956	2,792
$n_2>120$	6,635	4,605	3,782	3,319	3,017	2,802	2,639

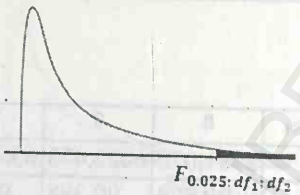
BUCHAREST UNIVERSITY OF ECONOMIC STUDIES LIBRARY

n_2/n_1	8	9	10	20	30	120	$n_1>120$
1	5981,070	6022,473	6055,847	6208,730	6260,649	6339,391	6365,864
2	99,374	99,388	99,399	99,449	99,466	99,491	99,499
3	27,489	27,345	27,229	26,690	26,505	26,221	26,125
4	14,799	14,659	14,546	14,020	13,838	13,558	13,463
5	10,289	10,158	10,051	9,553	9,379	9,112	9,020
6	8,102	7,976	7,874	7,396	7,229	6,969	6,880
7	6,840	6,719	6,620	6,155	5,992	5,737	5,650
8	6,029	5,911	5,814	5,359	5,198	4,946	4,859
9	5,467	5,351	5,257	4,808	4,649	4,398	4,311
10	5,057	4,942	4,849	4,405	4,247	3,996	3,909
11	4,744	4,632	4,539	4,099	3,941	3,690	3,602
12	4,499	4,388	4,296	3,858	3,701	3,449	3,361
13	4,302	4,191	4,100	3,665	3,507	3,255	3,165
14	4,140	4,030	3,939	3,505	3,348	3,094	3,004
15	4,004	3,895	3,805	3,372	3,214	2,959	2,868
16	3,890	3,780	3,691	3,259	3,101	2,845	2,753
17	3,791	3,682	3,593	3,162	3,003	2,746	2,653
18	3,705	3,597	3,508	3,077	2,919	2,660	2,566
19	3,631	3,523	3,434	3,003	2,844	2,584	2,489
20	3,564	3,457	3,368	2,938	2,778	2,517	2,421
21	3,506	3,398	3,310	2,880	2,720	2,457	2,360
22	3,453	3,346	3,258	2,827	2,667	2,403	2,305
23	3,406	3,299	3,211	2,781	2,620	2,354	2,256
24	3,363	3,256	3,168	2,738	2,577	2,310	2,211
25	3,324	3,217	3,129	2,699	2,538	2,270	2,169
26	3,288	3,182	3,094	2,664	2,503	2,233	2,131
27	3,256	3,149	3,062	2,632	2,470	2,198	2,097
28	3,226	3,120	3,032	2,602	2,440	2,167	2,064
29	3,198	3,092	3,005	2,574	2,412	2,138	2,034
30	3,173	3,067	2,979	2,549	2,386	2,111	2,006
40	2,993	2,888	2,801	2,369	2,203	1,917	1,805
60	2,823	2,718	2,632	2,198	2,028	1,726	1,601
120	2,663	2,559	2,472	2,035	1,860	1,533	1,381
$n_2>120$	2,511	2,407	2,321	1,878	1,696	1,325	1,000

Repartiția Fisher

$\alpha = 0{,}025$

$df_1 = n_1$, $df_2 = n_2$



n_2/n_1	1	2	3	4	5	6	7
1	647,789	799,500	864,163	899,583	921,847	937,111	948,216
2	38,5063	39,000	39,1655	39,2484	39,2982	39,3315	39,3552
3	17,4434	16,0441	15,4392	15,1010	14,8848	14,7347	14,6244
4	12,2179	10,6491	9,9792	9,6045	9,3645	9,1973	9,0741
5	10,0070	8,4336	7,7636	7,3879	7,1464	6,9777	6,8531
6	8,8131	7,2599	6,5988	6,2272	5,9876	5,8198	5,6955
7	8,0727	6,5415	5,8898	5,5226	5,2852	5,1186	4,9949
8	7,5709	6,0595	5,4160	5,0526	4,8173	4,6517	4,5286
9	7,2093	5,7147	5,0781	4,7181	4,4844	4,3197	4,1970
10	6,9367	5,4564	4,8256	4,4683	4,2361	4,0721	3,9498
11	6,7241	5,2559	4,6300	4,2751	4,0440	3,8807	3,7586
12	6,5538	5,0959	4,4742	4,1212	3,8911	3,7283	3,6065
13	6,4143	4,9653	4,3472	3,9959	3,7667	3,6043	3,4827
14	6,2979	4,8567	4,2417	3,8919	3,6634	3,5014	3,3799
15	6,1995	4,7650	4,1528	3,8043	3,5764	3,4147	3,2934
16	6,1151	4,6867	4,0768	3,7294	3,5021	3,3406	3,2194
17	6,0420	4,6189	4,0112	3,6648	3,4379	3,2767	3,1556
18	5,9781	4,5597	3,9539	3,6083	3,3820	3,2209	3,0999
19	5,9216	4,5075	3,9034	3,5587	3,3327	3,1718	3,0509
20	5,8715	4,4613	3,8587	3,5147	3,2891	3,1283	3,0074
21	5,8266	4,4199	3,8188	3,4754	3,2501	3,0895	2,9686
22	5,7863	4,3828	3,7829	3,4401	3,2151	3,0546	2,9338
23	5,7498	4,3492	3,7505	3,4083	3,1835	3,0232	2,9023
24	5,7166	4,3187	3,7211	3,3794	3,1548	2,9946	2,8738
25	5,6864	4,2909	3,6943	3,3530	3,1287	2,9685	2,8478
26	5,6586	4,2655	3,6697	3,3289	3,1048	2,9447	2,8240
27	5,6331	4,2421	3,6472	3,3067	3,0828	2,9228	2,8021
28	5,6096	4,2205	3,6264	3,2863	3,0626	2,9027	2,7820
29	5,5878	4,2006	3,6072	3,2674	3,0438	2,8840	2,7633
30	5,5675	4,1821	3,5894	3,2499	3,0265	2,8667	2,7460
40	5,4239	4,0510	3,4633	3,1261	2,9037	2,7444	2,6238
60	5,2856	3,9253	3,3425	3,0077	2,7863	2,6274	2,5068
120	5,1523	3,8046	3,2269	2,8943	2,6740	2,5154	2,3948
n>120	5,0239	3,6889	3,1161	2,7858	2,5665	2,4082	2,2875

n_2/n_1	8	9	10	20	30	120	$n_1>120$
1	956,656	963,284	968,627	993,102	1001,414	1014,020	1018,258
2	39,3730	39,3869	39,3980	39,4479	39,465	39,490	39,498
3	14,5399	14,4731	14,4189	14,1674	14,081	13,947	13,902
4	8,9796	8,9047	8,8439	8,5599	8,461	8,309	8,257
5	6,7572	6,6811	6,6192	6,3286	6,227	6,069	6,015
6	5,5996	5,5234	5,4613	5,1684	5,065	4,904	4,849
7	4,8993	4,8232	4,7611	4,4667	4,362	4,199	4,142
8	4,4333	4,3572	4,2951	3,9995	3,894	3,728	3,670
9	4,1020	4,0260	3,9639	3,6669	3,560	3,392	3,333
10	3,8549	3,7790	3,7168	3,4185	3,311	3,140	3,080
11	3,6638	3,5879	3,5257	3,2261	3,118	2,944	2,883
12	3,5118	3,4358	3,3736	3,0728	2,963	2,787	2,725
13	3,3880	3,3120	3,2497	2,9477	2,837	2,659	2,595
14	3,2853	3,2093	3,1469	2,8437	2,732	2,552	2,487
15	3,1987	3,1227	3,0602	2,7559	2,644	2,461	2,395
16	3,1248	3,0488	2,9862	2,6808	2,568	2,383	2,316
17	3,0610	2,9849	2,9222	2,6158	2,502	2,315	2,247
18	3,0053	2,9291	2,8664	2,5590	2,445	2,256	2,187
19	2,9563	2,8801	2,8172	2,5089	2,394	2,203	2,133
20	2,9128	2,8365	2,7737	2,4645	2,349	2,156	2,085
21	2,8740	2,7977	2,7348	2,4247	2,308	2,114	2,042
22	2,8392	2,7628	2,6998	2,3890	2,272	2,076	2,003
23	2,8077	2,7313	2,6682	2,3567	2,239	2,041	1,968
24	2,7791	2,7027	2,6396	2,3273	2,209	2,010	1,935
25	2,7531	2,6766	2,6135	2,3005	2,182	1,981	1,906
26	2,7293	2,6528	2,5896	2,2759	2,157	1,954	1,878
27	2,7074	2,6309	2,5676	2,2533	2,133	1,930	1,853
28	2,6872	2,6106	2,5473	2,2324	2,112	1,907	1,829
29	2,6686	2,5919	2,5286	2,2131	2,092	1,886	1,807
30	2,6513	2,5746	2,5112	2,1952	2,074	1,866	1,787
40	2,5289	2,4519	2,3882	2,0677	1,943	1,724	1,637
60	2,4117	2,3344	2,2702	1,9445	1,815	1,581	1,482
120	2,2994	2,2217	2,1570	1,8249	1,690	1,433	1,310
$n_2>120$	2,1918	2,1136	2,0483	1,7085	1,566	1,268	1,130

Repartiția Durbin-Watson

$\alpha = 0{,}05$; k reprezintă numărul de parametri din model

	k=2		*k=3*		*k=4*		*k=5*	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
7	0,700	1,356	0,467	1,896	-----	-----	-----	-----
8	0,763	1,332	0,559	1,777	0,367	2,287	-----	-----
9	0,824	1,320	0,629	1,699	0,455	2,128	0,296	2,588
10	0,879	1,320	0,697	1,641	0,525	2,016	0,376	2,414
11	0,927	1,324	0,758	1,604	0,595	1,928	0,444	2,283
12	0,971	1,331	0,812	1,579	0,658	1,864	0,512	2,177
13	1,010	1,340	0,861	1,562	0,715	1,816	0,574	2,094
14	1,045	1,350	0,905	1,551	0,767	1,779	0,632	2,030
15	1,077	1,361	0,946	1,543	0,814	1,750	0,685	1,977
16	1,106	1,371	0,982	1,539	0,857	1,728	0,734	1,935
17	1,133	1,381	1,015	1,536	0,897	1,710	0,779	1,900
18	1,158	1,391	1,046	1,535	0,933	1,696	0,820	1,872
19	1,180	1,401	1,074	1,536	0,967	1,685	0,859	1,848
20	1,201	1,411	1,100	1,537	0,998	1,676	0,894	1,828
21	1,221	1,420	1,125	1,538	1,026	1,669	0,927	1,812
22	1,239	1,429	1,147	1,541	1,053	1,664	0,958	1,797
23	1,257	1,437	1,168	1,543	1,078	1,660	0,986	1,785
24	1,273	1,446	1,188	1,546	1,101	1,656	1,013	1,775
25	1,288	1,454	1,206	1,550	1,123	1,654	1,038	1,767
26	1,302	1,461	1,224	1,553	1,143	1,652	1,062	1,759
27	1,316	1,469	1,240	1,556	1,162	1,651	1,084	1,753
28	1,328	1,476	1,255	1,560	1,181	1,650	1,104	1,747
29	1,341	1,483	1,270	1,563	1,198	1,650	1,124	1,743
30	1,352	1,489	1,284	1,567	1,214	1,650	1,143	1,739
31	1,363	1,496	1,297	1,570	1,229	1,650	1,160	1,735
32	1,373	1,502	1,309	1,574	1,244	1,650	1,177	1,732
33	1,383	1,508	1,321	1,577	1,258	1,651	1,193	1,730
34	1,393	1,514	1,333	1,580	1,271	1,652	1,208	1,728
35	1,402	1,519	1,343	1,584	1,283	1,653	1,222	1,726
36	1,411	1,525	1,354	1,587	1,295	1,654	1,236	1,724
37	1,419	1,530	1,364	1,590	1,307	1,655	1,249	1,723
38	1,427	1,535	1,373	1,594	1,318	1,656	1,261	1,722
39	1,435	1,540	1,382	1,597	1,328	1,658	1,273	1,722
40	1,442	1,544	1,391	1,600	1,338	1,659	1,285	1,721
45	1,475	1,566	1,430	1,615	1,383	1,666	1,336	1,720
50	1,503	1,585	1,462	1,628	1,421	1,674	1,378	1,721
55	1,528	1,601	1,490	1,641	1,452	1,681	1,414	1,724
60	1,549	1,616	1,514	1,652	1,480	1,689	1,444	1,727
65	1,567	1,629	1,536	1,662	1,503	1,696	1,471	1,731
70	1,583	1,641	1,554	1,672	1,525	1,703	1,494	1,735
75	1,598	1,652	1,571	1,680	1,543	1,709	1,515	1,739
80	1,611	1,662	1,586	1,688	1,560	1,715	1,534	1,743
85	1,624	1,671	1,600	1,696	1,575	1,721	1,550	1,747
90	1,635	1,679	1,612	1,703	1,589	1,726	1,566	1,751
95	1,645	1,687	1,623	1,709	1,602	1,732	1,579	1,755
100	1,654	1,694	1,634	1,715	1,613	1,736	1,592	1,758
150	1,720	1,747	1,706	1,760	1,693	1,774	1,679	1,788
200	1,758	1,779	1,748	1,789	1,738	1,799	1,728	1,809

BCU LIBRARY

În aceeaşi colecţie au apărut:

Management, Fundamente, studii de caz şi metode econometrice,
Ioan Ciobanu

Performanţa şi eficienţa activităţii bancare,
Alin Marius Andrieş

Moneda, creditul bancar şi ciclurile economice,
Jesús Huerta de Soto

Dezvoltarea aplicaţiilor orientate obiect pe platforma Java,
Cătălin Strîmbei

Diversitate culturală în management, O abordare interregională,
Angelica-Nicoleta Onea

Competitivitate industrială,
Petrică Corăbieru

Uniunea Europeană şi competiţia globală,
Roxana Paraschiv

Influenţa timpului asupra evaluării afacerilor, Analiză, Diagnostic, Evaluare,
Ioan Dumitrean

Costurile de tranzacţie – o abordare instituţionalistă,
Andreea-Oana Iacobuţă

Activitatea bancară şi integrarea monetară europeană,
Vasile Cocriş, Elena Sireteanu, Alin Marius Andrieş

Performanţă şi risc în afaceri, Concepte, metode, aplicaţii,
Silvia Petrescu

Auditul financiar, de la normele naţionale la standardele internaţionale,
Ionela-Corina Chersan

Performanţa în serviciile de sănătate publică,
Maria Viorica Bedrule-Grigoruţă (editor)

Diagnostic intercultural, Competitivitate organizaţională prin mixare culturală şi despre creşterea performanţei manageriale prin sinergie interculturală,
Dumitru Zaiţ

Schimburile comerciale internaţionale între teorie şi realitate,
Liviu-George Maha

Contabilitatea activelor şi datoriilor financiare la societăţile comerciale,
Maria Carmen Huian

Echilibrul extern în economia deschisă,
Lăcrămioara Juverdeanu

Pieţe de capital,
Carmen Corduneanu, Laura Raisa Miloş, Claudiu Boţoc

BCU IASI - UNIVERSITY LIBRARY

Impactul politicii de finanțare asupra dezvoltării durabile a întreprinderii, O abordare financiar-contabilă,
Mihai Carp

Finanțarea firmelor în economiile emergente,
Silviu Ursu

Dinamica economică în viziunea instituționalistă,
Oana-Ramona Socoliuc

Dinamica piețelor de capital emergente,
Delia-Elena Diaconașu

Regimurile cursului de schimb în contemporaneitate, Perspective neoliberale,
Iulian Ihnatov

Monetary, Banking and Financial Issues in Central and Eastern EU Member Countries: How Can Central and Eastern EU Members Overcome the Current Economic Crisis? (vol, I),
Angela Roman, Sorin Gabriel Anton (editors)

Monetary, Banking and Financial Issues in Central and Eastern EU Member Countries: How Can Central and Eastern EU Members Overcome the Current Economic Crisis? (Vol, II),
Angela Roman, Sorin Gabriel Anton (editors)

Monetary, Banking and Financial Issues in Central and Eastern EU Member Countries: How Can Central and Eastern EU Members Overcome the Current Economic Crisis? (Vol, III),
Angela Roman, Sorin Gabriel Anton (editors)

Diversitate culturală în management, O abordare interregională,
ed, a II-a, **Angelica-Nicoleta Onea**

Probleme actuale de economie publică,
Livia Baciu, Andreea Iacobuță, Alina Botezat, Mihaela Ifrim

Contabilitate în administrația publică,
Iuliana Georgescu, Leontina Păvăloaia

Gestiunea riscurilor financiare, Abordări teoretice și studii de caz,
ed, a II-a, **Sorin Gabriel Anton**

European Financial and Monetary Integration, Challenges of the Single Currency,
Angela Roman, Irina Bilan (editors)

Capitalul uman în noile democrații,
Cristian C, Popescu

Putere și globalizare,
Vol, I *Evoluția centrelor de putere economică până la sfârșitul secolului XX*
Cristian C, Popescu

Lecții de economie financiară, Cum și de ce investim,
Costică Mihai

Auditul intern, de la teorie la practică,
Ionela-Corina Chersan, Cristina-Ionela Precob

Economiile de aglomerare și competitivitatea regională,
Raluca Irina Clipa

Demografia României,
Dănuț-Vasile Jemna

BCU IASI CENTRAL UNIVERSITY LIBRARY

BCU IASI/CENTRAL UNIVERSITY LIBRARY

TIPARUL EXECUTAT LA
IMPRIMERIA EDITURII UNIVERSITĂȚII
„ALEXANDRU IOAN CUZA" DIN IAȘI
700109 Iași, Pînului 1A, tel./fax 0232 314947
Apărut: 2017
Comanda: 516

Informații și comenzi:
www.editura.uaic.ro
editura@uaic.ro

B.C.U. „M. EMINESCU" IAȘI

BCU IAȘI/CENTRAL UNIVERSITY LIBRARY

BCU IASI/CENTRAL UNIVERSITY LIBRARY

+ VALOARE

ECONOMIKON

Econometrie

cu aplicații în R

Ediția a IV-a revăzută și adăugită

DĂNUȚ-VASILE JEMNA

Una dintre cele mai importante inovații care au avut loc în domeniul economic, în secolul al XX-lea, a fost, cu siguranță, econometria. Folosind puterea cifrelor, cu ajutorul statisticii și matematicii, noii specialiști au schimbat modul de a gândi și de a înțelege realitatea economică. Plusul de cunoaștere oferit de evaluarea empirică a teoriilor economice, succesul cu care realitatea economică era explicată cu ajutorul modelelor matematice și statistice aveau să cucerească pe mulți, să producă polemici și dispute între economiști. O artă și o știință a utilizării datelor și teoriilor economice se dezvoltă cu rapiditate și își câștigă un loc special atât în spațiul academic, cât și în rândul practicienilor.

9 786067 143843
www.editura.uaic.ro

Editura Universității „Alexandru Ioan Cuza" Iași